VERSANT

JOHN BROCKMAN (ed.)

MINȚI POSIBILE

Douăzeci și cinci de perspective
asupra inteligenței artificiale

VELLANT
CĂRȚI ÎNSEMNATE.

JOHN BROCKMAN

Minţi Posibile

John Brockman este un impresar cultural a cărui carieră s-a intersectat cu domenii precum arta de avangardă, ştiinţa, publishingul, dezvoltarea de software sau Internetul. Este fondatorul platformei edge.org şi editorul seriei de cărţi Edge Question, care include titluri precum This Ideea is Brilliant, Know This, This Ideea Must Die, This Explains Everything, This Will Make You Smarter.

Redactor: Doinei Tronam
Corector: Duşa Udrea-Boborel
Copertă: Daniel Breanova
DTP: Răzvan Nasea
Coordonare editorială: Dom Someşan
Consultant editorial: Simona Radu-Răuţă
Concept şi identitate vizuală: RUSU + BORŢUN
Descrierea CIP a Bibliotecii Naţionale a României Minţi posibile: douăzeci şi cinci de perspective asupra inteligenţei artificiale / ed.: John Brockman; trad.: George Chiriţă.
— Bucureşti Editura Vellant, 2019 ISBN 978 – 606 – 980 – 067 – 6
I. Brockman, John (ed.)
II. Chiriţă, George (trad.)
004
Titlu original: *Possible Minds*: *Twenty-Five Ways of Looking at* Al Copyright © 22019 by John Brockman. All rights reserved.
© 22019 Editura Vellant
Editura Vellant
Splaiul Independenţei 319 Bucureşti, Sector 6 www.vdiant.ro

Pentru Einstein, Gertrude Stein, Wittgenstein şi Frankenstein

Mulţumiri

Multe mulţumiri lui Scott Moyers de la Penguin Press, pentru exuberanţa sa editorială, şi agentului meu, Max Brockman, pentru încurajările sale continue. Mulţumiri speciale, din nou, Sarei Lippincott, pentru atenţia acordată acestui manuscris.

Cuprins

Introducere
Despre promisiunile fi pericolele inteligenţei artificiale
Capitolul 1. Seth Lloyd:
Greşit, dar mai relevant decât oricând
Capitolul 2. Judea Pearl:
Limitările maşinilor de învăţare opace
Capitolul 3. Stuart Russell:
Scopul scris în maşină
Capitolul 4. George Dyson:
A treia lege
Capitolul 5. Daniel C. Dennett:
Ce putem face?
Capitolul 6. Rodney Brooks:
Debandada inumană în care ne-au băgat maşinile 84
Capitolul 7. Frank Wilezele
Unitatea inteligenţei
Capitolul 8. Max Tegmark:
Să aspirăm la mai mult decât doar să nu devenim depăşiţi
Capitolul 9. Jaan Tallinn:
Mesaje disidente
Capitolul 10. Steven Pinker:
Profeţia tehnologică şi puterea cauzală subapreciată a ideilor
Capitolul 11. David Deutsch:
Dincolo de recompensă şi pedeapsă
Capitolul 12. Tom Griffiths:
Folosirea artificială a fiinţelor umane
Capitolul 13. Anca Drăgan:
Integrarea omului în ecuaţia inteligenţei artificiale
Capitolul 14. Chris Anderson:
Coborârea pe panta unui gradient
Capitolul 15. David Kaiser:
Jnformaţiam pentru Wiener, pentru Shannon şi pentru noi
Capitolul 16. Neil Gershenfeld:
Scalarea
Capitolul 17. W. Daniel Hills:
Primele inteligenţe automate
Capitolul 18. Venki Ramakrishnan:
Vor deveni computerele stăpânii noştri
Capitolul 19. Alex „Sandy" Pentland:
Strategia umană
Capitolul 20. Hans Ulrich Obrist:

Transformarea invizibilului în vizibil: *artafaţă în faţă cu inteligenţa artificială*
Capitolul 21. Alison Gopnile
Inteligenţele artificiale versus copiii de patru ani
Capitolul 22. Peter Galison:
Algoritmicienii visează la obiectivitate
Capitolul 23. George M. Church:
Drepturile maşinilor
Capitolul 24. Caroline A. Jones:
Utilizarea artistică a fiinţelor cibernetice
Capitolul 25. Stephen Wolfram:
Inteligenţa artificială şi viitorul civilizaţiei

Introducere

Despre promisiunile şi pericolele inteligenţei artificiale

Inteligenţa artificială este povestea zilei de astăzi, povestea din spatele tuturor celorlalte poveşti, este a doua venire a lui Iisus şi Apocalipsa în acelaşi timp: inteligenţa artificială bună şi inteligenţa artificială malefică. Această carte a reieşit în urma unor conversaţii în desfăşurare cu câţiva gânditori importanţi, atât din lumea inteligenţei artificiale, cât şi din afara acesteia, despre ce este şi ce înseamnă inteligenţa artificială. Numită Proiectul Minţi Posibile, această conversaţie a început în luna septembrie 2016, printr-o întâlnire la Grace Mayflower Inn & Spa din Washington, Connecticut, cu câţiva dintre cei care au contribuit la această carte.

Ce a reieşit foarte repede din acea primă întâlnire a fost că entuziasmul şi teama din cultura mai largă ce înconjoară inteligenţa artificială acum au un corespondent în modul în care ideile lui Norbert Wiener cu privire la „cibernetică" şi-au făcut loc în cultură, în special în anii 1960, pe măsură ce artiştii au început să încorporeze gândirea despre noile tehnologii în activitatea lor. Am fost martor apropiat la impactul acelor idei; şi nu exagerez când spun că ele mi-au determinat traiectoria vieţii. Odată cu sosirea erei digitale, la începutul anilor 1970, oamenii au încetat să mai vorbească despre Wiener, dar astăzi această idee a ciberneticii a fost adoptată la scară atât de largă, încât este internalizată până la punctul în care nu mai are nevoie de nume. Este peste tot, este în aer şi este un loc potrivit de unde să începem.

Tehnologii noi = percepţii noi

Înainte de inteligenţa artificială, a existat cibernetica – ideea controlului automat, care se reglează singur, expusă în textul de referinţă al lui Norbert Wiener din 1948. Pot să-mi datez propria mea expunere serioasă la acest test în 1966, când compozitorul John Cage m-a invitat pe mine şi pe încă patru sau cinci artişti tineri să participăm împreună la o serie de serate – un seminar în desfăşurare despre mass-media, comunicaţii, artă, muzică şi filosofie, care se concentra pe interesul lui Cage în ceea ce priveşte ideile lui Wiener, Claude Shannon şi Marshall McLuhan, toţi având greutate în cercurile artistice din New York în care mă învârteam atunci. În special Cage preluase ideea lui McLuhan conform căreia, prin inventarea tehnologiilor electronice, am externalizat sistemul nostru nervos central – adică minţile noastre – şi acum trebuia să presupunem că

„există o singură minte, cea pe care o împărtăşim cu toţii”.

Ideile de acest fel începeau să fie interesante pentru artiştii cu care lucram la New York şi la Film-Makers, Cinematheque, unde eram director de programe artistice pentru o serie de producţii multimedia denumită New Cinema 1 (cunoscută şi sub numele Expanded Cinema Festival), sub auspiciile regizorului avangardist de film şi impresarului Jonas Mekas. Printre ei se numărau: Claes Oldenburg, Robert Rauschenberg, Andy Warhol şi Robert Whitman; artiştii cinetici: Charlotte Moorman şi n-am June Paik; artiştii momentului Allan Kaprow şi Carolee Schneemann; dansatoarea Trisha Brown; regizorii: Jack Smith, Stan Vanderbeek, Ed Emshwiller şi fraţii Kuchar; dramaturgul de avangardă Ken Dewey; poetul Gem Stern şi grupul USCO; muzicienii minimalişti La Monte Young şi Terry Riley; şi, prin Warhol, grupul muzical The Velvet Underground. Mulţi dintre aceşti oameni citeau Wiener şi cibernetica era în aer. La una dintre aceste serate, Cage a băgat mâna în valiza sa şi a scos un exemplar al cărţii *Cibernetica*, mi l-a înmânat şi a spus» Asta e pentru tine”.

În timpul festivalului, am primit un apel telefonic neaşteptat de la colegul lui Wiener, Arthur K. Solomon, coordonatorul masteratelor şi doctoratelor de la Harvard în biofizică. Wiener murise cu un an înainte şi ceilalţi colegi apropiaţi ai lui Wiener şi Solomon de la MIT şi Harvard citiseră despre Expanded Cinema Festival în *New York Times*, fiind intrigaţi de legătura cu cartea lui Wiener. Solomon m-a invitat să aduc câţiva artişti la Cambridge pentru a mă întâlni cu el şi cu un grup ce-i includea şi pe cercetătorul în domeniul comunicaţiilor senzoriale de la MIT, Walter Rosenblith, pe matematicianul de la Harvard Anthony Oettinger şi pe un inginer MIT Harold» Doc” Edgerton, inventatorul luminii stroboscopice.

Ca multe alte situaţii în care» arta întâlneşte ştiinţa” în care am fost implicat de atunci, evenimentul de două zile era un rateu informat: nave care se petrec în noapte. Dar m-am înhămat la acest proiect şi evenimentul a fost important din câteva puncte de vedere interesante – unul fiind faptul că ne-au dus să vedem Computerul. Computerele erau o raritate pe atunci; cel puţin niciunul dintre noi, cei care eram în vizită, nu văzuserăm vreodată unul. Am fost conduşi într-un spaţiu mare din campusul MIT, în mijlocul căruia se afla o» cameră rece” ridicată de la sol şi cu pereţi de sticlă, în care tehnicieni cu halate, eşarfe şi mănuşi albe erau ocupaţi cu colaţionarea unor cartele perforate care ieşeau dintr-o maşină enormă. Când m-am apropiat, aburul respiraţiei mele a aburit fereastra din camera rece. Ştergând fereastra, am văzut Computerul. M-am îndrăgostit.

Mai târziu, în toamna anului 1967, am mers în Menlo Park pentru a petrece puţin timp cu Stewart Brand, pe care l-am întâlnit în New York, în 1965, când era un membru-satelit al grupului de artişti USCO.

Acum, cu soţia sa Lois, matematiciană, pregătea pentru publicare prima ediţie a revistei *Whole Earth Catalog*. În timp ce Lois şi echipa făceau munca grea pentru ultimele mecanici ale *WECy* eu şi Stewart am stat împreună într-un colţ timp de două zile, citind, subliniind şi adnotând exemplarul din *Cibernetica* pe care mi-l dăduse Cage cu un an înainte şi dezbătând ideile lui Wiener.

Inspirat de acest set de idei, am început să dezvolt o temă, un fel de mantră, care mi-a ghidat încercările de atunci: „tehnologii noi înseamnă percepţii noi". Inspirat de teoreticianul comunicării Marshall McLuhan, designerul-arhitect Buckminster Fuller, futuristul John McHale şi antropologii culturali Edward T. „Ned" Hall şi Edmund Carpenter, am început să citesc cu nesaţ despre teoria informaţiei, cibernetică şi teoria sistemelor. McLuhan mi-a sugerat să citesc cartea biologului J.Z. Young *Doubt and Certainty în Science*, în care spune că noi creăm instrumente şi ne modelăm prin utilizarea acestora. Celălalt text pe care l-a recomandat era lucrarea din 1949 a lui Warren Weaver şi Claude Shannon intitulată *Contribuţii recente la teoria matematică a comunicării*, care începe aşa: „Cuvântul *comunicare* va fi utilizat aici cu sens foarte larg, pentru a include toate procedurile prin care o minte o poate afecta pe alta. Bineînţeles, acest lucru nu implică numai limbajul scris şi oral, ci şi muzica, pictura, teatrul, baletul şi, de fapt, tot comportamentul uman".

Cine ar fi ştiut că în două decenii de la acel moment vom începe să recunoaştem creierul ca pe un computer? Şi, în următoarele două decenii, pe măsură ce am integrat computerele noastre pe internet, că va urma să realizăm despre creierul nostru că nu este un computer, ci o reţea de computere? Cu siguranţă nu Wiener, specialist în circuitele cu *feedback* analogice concepute pentru a controla maşinile, nici artiştii, cu atât mai puţin eu.

„Trebuie să încetăm să sărutăm biciul ce ne biciuieşte".

La doi ani după *Ciberneticay* în 1950, Norbert Wiener a publicat cartea *The Human Use of Human Beings*[1] o poveste mai profundă, în care îşi exprima îngrijorările despre exploatarea comercială scăpată de sub control şi alte consecinţe neprevăzute ale noilor tehnologii de control. Nu am citit *The Human Use of Human Beings* până în primăvara anului 2016, când am găsit un exemplar, ediţie princeps, care stătea în biblioteca mea lângă *Cibernetica*. Ce m-a şocat a fost să realizez cât de profetic a putut să fie Wiener în 1950 despre ce se întâmplă în zilele noastre. Deşi prima ediţie a fost un *bestseller* major – şi, într-adevăr, a iniţiat o conversaţie importantă –, sub presiunea colegilor săi de breaslă, Wiener a scos o ediţie revizuită şi mai temperată în 1954, din care capitolul de concluzii iniţial, „Vocile rigidităţii", lipseşte.

Istoricul ştiinţei George Dyson evidenţiază faptul că, în această ediţie princeps de mult uitată, Wiener a preconizat posibilitatea „unui fascism nou şi ameninţător, dependent de *machine a gouverner*”:

> *Nicio elită nu a scăpat de critica sa, de la marxişti şi iezuiţi („tot catolicismul este esenţialmente o religie totalitară”) până la FBI („marii noştri prinţi negustori au analizat tehnica de propagandă a ruşilor şi au constatat că este bună”) şi mogulii care îşi oferă sprijinul în „a face capitalismul american şi a cincea libertate a comerciantului supreme în lume”. Oamenii de ştiinţă… au fost supuşi aceleiaşi cercetări meticuloase la care a fost supusă şi Biserica: „Într-adevăr, directorii marilor laboratoare seamănă foarte mult cu episcopii, cu asocierile lor cu persoane influente din toate palierele vieţii şi pericolele la care se supun din cauza păcatelor carnale, a mândriei şi a lăcomiei de putere”.*

Ieremiada nu a mers bine pentru Wiener. În cuvintele lui Dyson:

> *Acele semnale de alarmă au fost ignorate la vremea aceea nu pentru că Wiener se înşela în ceea ce priveşte tehnica de calcul digitală, ci pentru că existau ameninţări mai mari când şi-a terminat manuscrisul în toamna anului 1949. Wiener nu avea nimic împotriva tehnicii de calcul digitale, dar se opunea cu tărie armelor nucleare şi a refuzat să se alăture celor care construiau computere digitale pentru a avansa cercetările pentru bomba cu hidrogen, care era de o mie de ori mai puternică.*

Deoarece prima ediţie a cărţii *The Human Use of Human Beings* este acum ieşită de pe piaţă, nu mai avem acces la *cri-de-coeur-vi* lui Wiener, mai relevant astăzi decât atunci când l-a scris, acum şaizeci şi opt de ani: „Trebuie să încetăm să sărutăm biciul ce ne biciuieşte”.

Mintea, gândirea, inteligenţa

Printre motivele pentru care nu auzim prea multe despre cibernetică în zilele noastre, două sunt centrale: în primul rând, deşi *The Human Use of Human Beings* a fost considerată o carte importantă la vremea sa, era contrară aspiraţiilor multor colegi ai lui Wiener, printre care şi John von Neumann şi Claude Shannon, care erau interesaţi de comercializarea noilor tehnologii. În al doilea rând, pionierul computerelor, John McCarthy, nu-l agrea pe Wiener şi refuza să folosească termenul acestuia de „cibernetică”. McCarthy, la rândul său, a născocit sintagma de „inteligenţă artificială” şi a devenit un „părinte fondator” al acestui domeniu.

După cum îmi explica Judea Pearl, care, în anii 1980, a introdus o abordare nouă a inteligenţei artificiale, denumită reţelele bayesiene:

Ceea ce a creat Wiener a fost entuziasmul de a crede că într-o bună zi vom construi o maşină inteligentă. Nu era un informatician. Vorbea despre feedbacky despre comunicare, despre analogic. Metafora prin care se exprima era un circuit cu feedback, domeniu în care era expert. Când a debutat era digitală, la începutul anilor 1960, oamenii voiau să vorbească despre programare, despre coduri, despre funcţii computaţionale, despre memoria pe termen scurt, memoria pe termen lung – metafore informatice care musteau de sens. Wiener nu făcea parte din acest val şi nu a ajuns la noua generaţie ce se născuse cu ideile sale. Metaforele sale erau prea vechi, passé. Erau deja disponibile mijloace noi care erau pregătite să captiveze imaginaţia umană. Pe la începutul anilor 1970, oamenii nu mai vorbeau despre Wiener.

Un factor critic care lipsea din viziunea lui Wiener era elementul cognitiv: mintea, gândirea şi inteligenţa. Încă din 1942, la prima întâlnire dintr-o serie de întâlniri interdisciplinare despre controlul sistemelor complexe care vor fi cunoscute sub numele de Conferinţele Macy, cercetători de top argumentau includerea elementului cognitiv în conversaţie. În timp ce von Neumann, Shannon şi Wiener erau concentraţi pe sistemele de control şi comunicare ale sistemelor observate, Warren McCullough voia să includă şi mintea. S-a orientat către antropologii culturali Gregory Bateson şi Margaret Mead, pentru a face legătura cu ştiinţele sociale. Bateson, în special, vorbea din ce în ce mai mult despre tipare şi procese sau despre „tiparul care conectează". A solicitat un nou tip de ecologie a sistemelor, în care organismele şi mediul în care trăiesc sunt unul şi acelaşi şi ar trebui considerate un singur circuit. La începutul anilor 1970, cibernetica sistemelor observate – cibernetica de rangul întâi – s-a deplasat către cibernetica sistemelor observatoare – cibernetica de rangul al doilea sau „cibernetica ciberneticii", termen introdus de Heinz von Foerster, care s-a alăturat conferinţelor Macy la jumătatea anilor 1950 şi a condus o nouă mişcare.

Cibernetica, în loc să dispară, se metaboliza în *tot* aşa că nu o mai vedeam ca pe o nouă disciplină separată şi distinctă. Şi acolo a rămas, ascunsă în văzul tuturor.

„Sceneta actorilor principali".

Propriile mele scrieri pe această temă la momentul respectiv erau pe radarul oamenilor care se ocupau de cibernetica de rangul al doilea, printre care Heinz von Foerster, dar şi John Lilly şi Alan Watts, care erau coorganizatorii a ceva numit Conferinţele AUM, acronim pentru „the American University of Masters", care a avut loc în Big Sur, în 1973, o adunare de filosofi, psihologi şi oameni de ştiinţă,

fiecare personalitate trebuind să ţină o prelegere despre propria sa activitate în termenii relaţiei acesteia cu ideile matematicianului britanic G. Spencer-Brown, prezentate în cartea sa *Laws of Form*. Am fost puţin nedumerit când am primit invitaţia – o invitaţie foarte târzie –, despre care spuneau că se bazează pe interesul lor în ideile prezentate de mine într-o carte cu titlul *Afterwords*, idei care erau pe lungimea lor de undă. Am profitat bucuros de ocazie, motivul principal fiind că vorbitorul principal era nimeni altul decât Richard Feynman. Ador să-mi petrec timpul cu fizicieni, deoarece se gândesc la univers, adică la tot. Şi niciun fizician nu avea o reputaţie de a fi mai articulat decât Feynman. Abia aşteptam să-l cunosc. Am acceptat. Acestea fiind spuse, nu sunt un om de ştiinţă şi nu m-am gândit niciodată să mă urc pe o scenă şi să ţin vreo „prelegere" de orice fel, cu atât mai puţin un comentariu despre o teorie matematică obscură în faţa unui grup identificat ca fiind format dintre cei mai interesanţi gânditori din lume. Abia după ce am sosit în Big Sur, am aflat motivul invitaţiei mele întârziate. „Când va vorbi Feynman?" am întrebat eu la recepţie. „A, nu ţi-a spus Alan Watts? Richard este bolnav şi a fost internat. Eşti înlocuitorul lui. Şi, apropo, care estetidul discursului tău inaugural?"

Am încercat să mă fac invizibil câteva zile. Alan Watts, dându-şi seama că evit podiumul, m-a trezit într-o noapte, la trei dimineaţa, ciocănind la uşă. Am deschis uşa şi l-am văzut stând în faţa mea într-o robă de călugăr, cu gluga acoperindu-i o mare parte din faţă. Avea braţele întinse, o lanternă într-o mână şi un whisky Magnum în cealaltă. John", a spus el pe o voce groasă şi cu un accent britanic aristocratic, „eşti un impostor. Şi, John", a continuat „eu sunt un impostor. Dar, John, eu sunt un impostor *adevăraţi*".

A doua zi, am ţinut discursul, intitulat „Einstein, Gertrude Stein, Wittgenstein şi Frankenstein". Einstein, revoluţia din fizica secolului XX. Gertrude Stein, prima scriitoare care şi-a integrat în lucrări ideea unui univers nedeterminat şi discontinuu. Cuvintele nu reprezentau nici caracterul, nici activitatea: un trandafir este un trandafir este un trandafir, iar un univers este un univers este un univers. Wittgenstein: lumea ca limite ale limbajului. „Limitele limbajului meu înseamnă limitele lumii mele". Sfârşitul distincţiei dintre observator şi observat. Frankenstein: cibernetica, inteligenţa artificială, robotica, toţi eseiştii din acest volum.

Prelegerea a avut consecinţe neprevăzute. Printre participanţii la Conferinţele AUM erau mai mulţi autori de bestselleruri ajunse numărul unu în *New York Timesy* însă nimeni de acolo nu avea un agent literar. Şi mi-am dat seama că toată lumea scria un gen de carte nedenumit şi nerecunoscut de editurile din New York. Deoarece aveam un masterat în administrarea afacerilor de la Columbia

Business School şi o serie de succese relative în afaceri, am fost hărţuit să devin agent, iniţial pentru Gregory Bateson şi John Lilly, ale căror cărţi le-am vândut rapid şi pentru sume care mi-au captat atenţia, începându-mi astfel cariera de agent literar.

Nu am apucat să-l întâlnesc niciodată pe Richard Feynman.

Lungile ierni ale inteligenţei artificiale

Această nouă carieră m-a pus în legătură apropiată cu majoritatea pionierilor din domeniu şi, de-a lungul deceniilor, am plutit cu ei pe valuri de entuziasm şi în văi de dezamăgire. La începutul anilor 1980, guvernul japonez a făcut un efort naţional de a avansa inteligenţa artificială. L-au numit a cincea generaţie; obiectivul lor era să schimbe arhitectura tehnicii de calcul trecând de „problema limitării a lui von Neumann" şi creând un computer care lucrează masiv în paralel. Făcând acest lucru, sperau să-şi dezvolte economia şi să devină o putere mondială dominantă în domeniu. În 1983, liderul consorţiului celei de-a Cincea Generaţii Japoneze a venit în New York pentru o întâlnire organizată de Heinz Pagels, preşedintele Academiei de Ştiinţe din New York. Aveam un loc la masă, alături de liderii primei generaţii Marvin Minsky şi John McCarthy; de cei ai celei de-a doua generaţii, Edward Feigenbaum şi Roger Schank; şi de Joseph Traub, directorul Consorţiului Naţional de Supercomputere.

În 1981, cu ajutorul lui Heinz, am înfiinţat The Reality Club (precursorul organizaţiei nonprofit Edge.org), ale cărei întâlniri interdisciplinare iniţiale aveau loc în sala de consiliu a Academiei de Ştiinţe din New York. Heinz lucra la cartea sa *The Dreams of Reason: The Computer and the rise of the Science of Complexity*, pe care o considera a fi o agendă de cercetări pentru ştiinţă în anii 1990.

La întâlnirile Reality Club, am cunoscut doi cercetători tineri care urmau să joace roluri-cheie în revoluţionarea informaticii. La MIT, la sfârşitul anilor 1970, Danny Hillis a dezvoltat algoritmii care au făcut posibil computerul masiv în paralel. În 1983, compania sa, Thinking Machines, a construit cel mai rapid supercomputer din lume utilizând arhitectura în paralel. „Maşina sa de conexiuni" reflecta îndeaproape modul în care funcţionează mintea omenească. Seth Lloyd, de la Universitatea Rockefeller, desfăşură o activitate foarte importantă în domeniile tehnicii de calcul cuantice şi comunicaţiilor cuantice, propunând inclusiv primul design fezabil din punct de vedere tehnologic pentru un computer cuantic.

Iar japonezii? Incursiunea lor în domeniul inteligenţei artificiale a eşuat şi a fost urmată de douăzeci de ani de creştere ecoy nomică anemică. Însă cercetătorii importanţi din SUA au luat acest program foarte în serios. Iar Feigenbaum, care era informaticianul cel mai cunoscut al momentului, a făcut echipă cu Pamela McCorduck pentru

a scrie o carte despre aceste dezvoltări. *The Fifth Generation*: *Artificial Intelligence and Japaris Computer*

Challenge to the World a fost publicată în 1983. Aveam un nume de cod pentru proiect: „Vine! Vine!" Dar nu a venit; a plecat.

Din acel moment, am lucrat cu cercetători în aproape orice varietate a inteligenţei artificiale şi complexităţii, printre care: Rodney Brooks, Hans Moravec, John Archibald Wheeler, Benoît Mandelbrot, John Henry Holland, Danny Hillis, Freeman Dyson, Chris Langton, J. Doyne Farmer, Geoffrey West, Stuart Russell şi Judea Pearl.

Un sistem emergent dinamic în desfăşurare

De la întâlnirea iniţială din Washington, Connecticut, până în prezent, am organizat mai multe serate şi discuţii în Londra şi Cambridge, Massachusetts, precum şi un eveniment public la Primăria Londrei. Printre invitaţi s-au numărat oameni de ştiinţă distinşi, istorici ai ştiinţelor şi teoreticieni ai comunicării, toţi gândindu-se serios întreaga lor carieră la problemele inteligenţei artificiale.

Am comandat eseuri unei plaje largi de participanţi, cu sau fără referire la Wiener (lăsând acest lucru la latitudinea fiecărui participant). În cele din urmă, douăzeci şi cinci de oameni au scris eseuri, fiecare în parte fiind îngrijorat de ce se întâmplă astăzi în era inteligenţei artificiale. *Minţi posibile* nu este cartea *mea*, mai degrabă este cartea *noastră*: Seth Lloyd, Judea Pearl, Stuart Russell, George Dyson, Daniel C. Dennett, Rodney Brooks, Frank Wilczek, Max Tegmark, Jaan Tallinn, Steven Pinker, David Deutsch, Tom Griffiths, Anca Dragan, Chris Anderson, David Kaiser, Neil Gershenfeld, W. Daniel Hillis, Venki Ramakrishnan, Alex „Sandy" Pentland, Hans Ulrich Obrist, Alison Gopnik, Peter Galison, George M. Church, Caroline A. Jones şi Stephen Wolfram.

Văd Proiectul Minţi Posibile ca pe un sistem emergent dinamic în desfăşurare, o prezentare a ideilor unei comunităţi de gânditori sofisticaţi, care îşi pun la treabă experienţa şi erudiţia în provocarea poveştii prevalente despre inteligenţa artificială digitală, comunicându-şi gândurile unul altuia. Ţinta este să prezentăm un mozaic de perspective care va ajuta să clarificăm acest domeniu emergent care se dezvoltă rapid.

Le-am cerut eseiştilor să ia în considerare:

a. *Poemul de tip zen „Treisprezece chipuri de a privi o mierlă"*[2]*, de Wallace Stevens, despre care insista că „nu a fost intenţionat a fi o colecţie de epigrame sau de idei, ci de senzaţii". Este un exerciţiu de „perspectivism", constând în secţiuni mici şi separate, fiecare menţionând mierlele într-un anumit fel. Poemul este despre propria sa imaginaţie; priveşte lucrurile de*

care se ocupă el.

b. *Parabola orbilor şi elefantului. La fel ca elefantul, inteligenţa artificială este o temă prea mare pentru o singură perspectivă, fără a mai pune la socoteală faptul că nu există doi oameni care să privească lucrurile la fel.*

Ce vrem să facă această carte? Stewart Brand a observat că „revizitarea gândirii pionierilor este întotdeauna utilă. Şi oferă o perspectivă amplă, care invită la gândire decenii şi chiar secole despre un subiect. Toate discuţiile contemporane sunt sortite să îmbătrânească urât şi imediat fără perspectiva pe termen lung".

Danny Hillis vrea ca oamenii din domeniul inteligenţei artificiale să îşi dea seama cum au fost programaţi de cartea lui Wiener. „Executaţi planul său", spune el, „şi pur şi simplu nu vă daţi seama".

Dan Dennett ar dori să „îl lase pe Wiener să se ridice ca o fantomă la banchet. Gândeşte-te la el ca la o sursă de vigoare hibridă, o sursă de idei fără odihnă care scutură de umeri mentalitatea leneşă".

Neil Gershenfeld argumentează că „educaţia corectoare sub acoperire pentru oamenii care trăiesc cu trăsăturile de personalitate «Big Five» ar fi un rezultat important al cărţii".

Freeman Dyson, unul dintre puţinii oameni în viaţă care l-au cunoscut pe Wiener, observă că „*The Human Use of Human Beings* este una dintre cele mai bune cărţi scrise vreodată. Wiener a nimerit aproape totul. Sunt interesat să văd ce va face cu aceste informaţii hoarda ta de vrăjitori".

Povestea în desfăşurare a inteligenţei artificiale

Lucrurile s-au schimbat şi rămân la fel. Acum inteligenţa artificială este peste tot. Avem internetul. Avem telefoanele noastre inteligente. Fondatorii companiilor dominante – companiile care deţin „biciul ce ne biciuieşte" – au venituri nete de 65 de miliarde de dolari, 90 de miliarde de dolari, 130 de miliarde de dolari. Persoane publice influente precum Elon Musk, Nick Bostrom, Martin Rees, Eliezer Yudkowsky şi regretatul Stephen Hawking au emis avertismente sumbre despre inteligenţa artificială, care au avut ca rezultat ascensiunea unor institute bine finanţate care au sarcina de a promova „inteligenţa artificială bună". Dar vom putea noi, ca specie, să controlăm o inteligenţă artificială pe deplin realizată, nesupravegheată şi care se îmbunătăţeşte automat? Avertizările şi dojanele lui Wiener din *The Human Use of Human Beings* sunt acum reale şi trebuie analizate cu ochi proaspeţi de cercetătorii din prima linie a revoluţiei inteligenţei artificiale. Iată-l pe Dyson din nou:

Wiener a respins din ce în ce mai mult ideea „veneratorilor de

gadgeturi", al căror egoism corporatist a creat „motive de automatizare care trec dincolo de o curiozitate legitimă, fiind păcătoase în sine". A ştiut că pericolul nu este reprezentat de maşinile care se vor asemui din ce în ce mai mult cu oamenii, ci de oamenii trataţi ca nişte maşini. „Lumea viitorului va fi o luptă din ce în ce mai aprigă împotriva limitărilor inteligenţei noastre", avertiza el in God & Golem, Inc., carte publicată în 1964, în anul morţii sale, „nu un hamac confortabil, în care ne putem tolăni, în timp ce sclavii noştri roboţi ne vor îndeplini toate capriciile".

Este timpul să examinăm povestea în desfăşurare a inteligenţei artificiale, identificându-i pe membrii principali ai comunităţii *mainstream*, dar şi pe disidenţi şi prezentându-le contrapoveştile în propriile lor cuvinte.

Eseurile care urmează constituie astfel o actualizare care era necesară de mult timp în domeniu.

John Brockman
New York, 2019

Capitolul 1
Greşit, dar mai relevant decât oricând

Seth Lloyd

Seth Lloyd este fizician teoretician la MIT, profesor N-am P. Suh în Departamentul de Inginerie Mecanică şi profesor colaborator la Institutul din Santa Fé.

L-am cunoscut pe Seth Lloyd la sfârşitul anilor 1980, când moduri noi de gândire erau peste tot: importanţa principiilor biologice de organizare, perspectiva informatică asupra matematicii şi proceselor fizice, accentul pus pe reţelele construite în paralel, importanţa dinamicii neliniare, noua înţelegere a haosului, ideile conecţioniste, reţele neuronale şi prelucrarea distributivă paralelă. Progresele făcute în domeniul tehnicii de calcul în acea perioadă ne-au oferit o nouă perspectivă asupra cunoaşterii.

Lui Seth îi place să se numească „mecanic cuantic". Este cunoscut internaţional pentru activitatea sa în domeniul tehnicii de calcul cuantice, care încearcă să controleze proprietăţile exotice ale teoriei cuantice, cum ar fi superpoziţia şi entanglementul, pentru a rezolva probleme care ar necesita câteva vieţi de om dacă ar fi rezolvate de computerele clasice.

În eseul care urmează, urmăreşte istoria informaţiei de la observaţiile profetice ale lui Norbert Wiener până la predicţiile unei „singularităţi" tehnologice, despre care unii ar vrea să credem că va înlocui specia umană. Observaţiile sale privind metoda recentă de programare cunoscută sub numele de învăţare profundă propun să avem un set mai modest de aşteptări; spune că, în ciuda progreselor uriaşe pe care le face inteligenţa artificială, „roboţii încă nu se pot lega la şireturi".

Este dificil pentru mine să vorbesc despre Seth fără a face referire la relaţia sa cu prietenul şi profesorul său, regretatul fizician teoretician Heinz Pagels de la Universitatea Rockefeller. Studentul şi profesorul au avut, fiecare, un efect profund asupra ideilor celuilalt.

În vara anului 1988, i-am vizitat pe Heinz şi Seth la Centrul de Fizică din Aspen. Activitatea lor comună pe tema complexităţii a fost prezentată în numărul curent din Scientific American, au fost exuberanţi. Asta se întâmpla cu numai două săptămâni înainte de moartea tragică a lui Heinz într-un accident montan, în timp ce cobora de pe Pyramid Peak cu Seth. Vorbeau despre tehnica de

calcul cuantică.

The Human Use of Human Beings, popularizarea de către Norbert Wiener din anii 1950 a influentei sale cărţi *Cibernetica*: *sau controlul şi comunicarea la animale şi maşini* (1948) [3], investighează interdependenţa dintre oameni şi maşini într-o lume în care maşinile devin din ce în ce mai performante din punctul de vedere al puterii de calcul. Este o remarcabilă carte vizionară… şi remarcabil de greşită. Scrisă în toiul Războiului Rece, conţine o avertizare înfiorătoare despre pericolele organizaţiilor şi societăţilor totalitare, dar şi ale democraţiei atunci când încearcă să combată totalitarismul cu arme totalitare.

Cibernetica lui Wiener analiza cu minuţiozitate ştiinţifică procesul de comandă prin *feedback* („cibernetica", din termenul grecesc antic pentru „timonier", este baza etimologică pentru cuvântul nostru[4] „governor", termenul pe care l-a folosit James Watt pentru a-şi denumi dispozitivul revoluţionar de comandă automată cu feedback care a transformat folosirea motoarelor cu abur.) Deoarece se ocupa de probleme legate de comandă, Wiener vedea lumea ca pe un set complex de cicluri *feedback* interdependente, în care senzori, semnale şi dispozitive de acţionare interacţionează prin intermediul unui schimb complex de semnale şi informaţii. Aplicaţiile inginereşti ale *ciberneticii* au avut o influenţă şi o eficienţă colosale, dând naştere rachetelor, roboţilor, liniilor automate de asamblare şi unei serii întregi de tehnici inginereşti de precizie – în alte cuvinte, au pus bazele societăţii industriale contemporane.

Însă Wiener a avut ambiţii mai mari pentru conceptele cibernetice şi, în *The Human Use of Human Beings*, îşi prezintă ideile cu privire la aplicarea conceptelor sale în domenii diverse precum Demonul lui Maxwell[5], limbajul omenesc, creier, metabolismul insectelor, sistemul legal, rolul inovaţiei tehnologice în procesul de guvernare şi chiar în domeniul religiei. Aceste aplicaţii mai largi ale ciberneticii au fost o nereuşită aproape fără echivoc. Promovată viguros de la sfârşitul anilor 1940 până la începutul anilor 1960 – similar cu promovarea tehnologiilor informaţiei şi comunicaţiilor care a condus la prăbuşirea dot-com din anii 2000 – 2001 –, cibernetica oferea sateliţi şi sisteme de comutatoare pentru telefonie, însă genera puţine dezvoltări utile sau chiar deloc în organizarea socială sau a societăţii în ansamblu.

Însă, după şaptezeci de ani, *The Human Use of Human Beings* ne poate învăţa mai multe pe noi oamenii decât o putea face la început. Poate că cea mai remarcabilă caracteristică a cărţii este că prezintă un număr mare de subiecte cu privire la interacţiunea dintre oameni şi maşini, care au încă o relevanţă considerabilă. Pe un ton sumbru, cartea face mai multe predicţii despre dezastrele care urmau să apară în a doua jumătate a secolului XX, multe dintre ele fiind aproape

identice cu predicţiile făcute în zilele noastre despre a doua jumătate a secolului al XXI-lea.

De exemplu, Wiener a prevăzut un moment în viitorul apropiat de anul 1950 în care oamenii ar urma să predea controlul societăţii unei inteligenţe artificiale cibernetice, care va provoca apoi numai dezastre umanităţii. Automatizarea producţiei, preconiza Wiener, urma să creeze avantaje mari în ceea ce priveşte productivitatea, dar şi să lase mulţi oameni fără locuri de muncă – o secvenţă de evenimente care a avut într-adevăr loc în deceniile care au urmat. Dacă societatea nu urma să găsească ocupaţii productive pentru aceşti muncitori rămaşi fără locuri de muncă – avertiza Wiener vor apărea inevitabil revolte.

Dar Wiener nu a prevăzut dezvoltări tehnologice cruciale. Ca aproape toţi tehnologii anilor 1950, nu a reuşit să prevadă revoluţia computerelor. Computerele, credea el, se vor ieftini de la sute de mii de dolari (cât costau în anii 1950) la zeci de mii de dolari; nici el, nici colegii săi de breaslă nu au anticipat explozia uriaşă a puterii computerelor care va urma dezvoltării tranzistorului şi circuitelor integrate. În cele din urmă, din cauza accentului pe care-l punea pe control, Wiener nu putea să conceapă o lume tehnologică în care inovaţia şi autoorganizarea mai degrabă se ridică de la fund ca bulele în pahar decât să fie impuse de deasupra.

Concentrându-se pe răul totalitarismului (politic, ştiinţific şi religios), Wiener vedea lumea într-o lumină profund pesimistă. Cartea sa ne avertiza de catastrofa care ne paşte dacă nu ne corectăm comportamentul, rapid. Lumea contemporană compusă din fiinţe umane şi maşini, la peste jumătate de secol după publicarea cărţii sale, este mult mai complexă şi bogată, conţinând o varietate mult mai mare de sisteme politice, sociale şi ştiinţifice decât a putut el să conceapă. Însă avertizările pe care ni le-a prezentat cu privire la ce se poate întâmpla dacă greşim – de exemplu, controlul întregului internet de un regim totalitar global – rămân la fel de relevante şi apăsătoare astăzi pe cât erau în anii 1950.

Ce a intuit corect Wiener

Cele mai cunoscute lucrări matematice ale lui Wiener se concentrau pe probleme de analiză a semnalului şi efectele zgomotului. În timpul celui de-al Doilea Război Mondial, a dezvoltat tehnici de ţintire pentru avioanele de luptă, concepând modele ce puteau prognoza traiectoria viitoare a unui avion extrapolând comportamentul său anterior. În *Cibernetica* şi în *The Human Use of Human Beings*, Wiener observă că acest comportament trecut include elementele normale şi anormale ale comportamentului pilotului uman, prin urmare un dispozitiv mecanizat putea prognoza comportamentul oamenilor. Ca Alan Turing, al cărui test Turing sugera că maşinile de

calcul pot oferi răspunsuri la întrebări ce nu pot fi distinse de răspunsurile omeneşti, Wiener era fascinat de noţiunea capturării comportamentului uman în descrieri matematice. În anii 1940, şi-a aplicat cunoştinţele sale despre ciclurile de control şi *feedback* pe *feedbackul* neuromuscular din sistemele vii şi a fost responsabil pentru aducerea lui Warren McCulloch şi Walter Pitts la MIT, unde au început activitatea lor de pionierat în studiul reţelelor neuronale artificiale.

Observaţia principală a lui Wiener a fost că lumea trebuie înţeleasă ca informaţie. Sistemele complexe, cum sunt, de exemplu, organismele, creierele şi societăţile umane, constau în cicluri de *feedback* interconectate, în care semnalele schimbate între subsisteme rezultă în comportamente complexe, dar stabile. Când ciclurile de *feedback* nu mai funcţionează, sistemul devine instabil. A construit o imagine convingătoare a modului de funcţionare a funcţiei sistemelor biologice complexe, o imagine care este acceptată aproape universal şi astăzi.

Viziunea lui Wiener asupra informaţiei drept cantitate centrală în guvernarea comportamentului sistemelor complexe a fost remarcabilă la momentul respectiv. În zilele noastre, când automobilele şi frigiderele sunt înţesate de microprocesoare şi o mare parte din societatea omenească se învârte în jurul computerelor şi celularelor conectate prin internet, pare prozaic să insistăm pe caracterul central al informaţiei, tehnicii de calcul şi comunicării. Pe vremea lui Wiener însă primele computere digitale abia fuseseră inventate, iar internetul nu era nici măcar o sclipire în colţul ochiului unui tehnolog.

Concepţia puternică a lui Wiener conform căreia nu doar sistemele complexe proiectate, ci toate sistemele complexe au la bază cicluri de semnale şi calcule a condus la contribuţii majore la dezvoltarea sistemelor complexe făcute de om. Metodele pe care el şi alţii le-au dezvoltat pentru controlul rachetelor, de exemplu, au fost folosite mai târziu la construirea rachetei selenare Saturn V, una dintre cele mai mari realizări inginereşti din secolul XX. În special, aplicarea conceptelor cibernetice ale lui Wiener în domeniile creierului şi percepţiei computerizate sunt precursorii direcţi ai circuitelor de învăţare profundă bazate pe reţele neuronale şi ai inteligenţei artificiale înseşi. Dar dezvoltările curente din aceste domenii au deviat de la viziunea sa, iar dezvoltarea lor viitoare poate afecta temeinic atât folosirea omului de către om, cât şi a maşinilor de către om.

Ce a intuit greşit Wiener

Exact în ceea ce priveşte extinderea ideii cibernetice la fiinţele umane concepţiile lui Wiener şi-au ratat ţinta. Lăsând la o parte deocamdată cugetările sale privind limbajul, legea şi societatea omenească, priviţi inovaţiile mai umile, dar cu potenţial utilitar, pe

care le considera iminente în 1950. Wiener observă că protezele ar fi mult mai eficiente dacă purtătorii lor ar putea comunica direct cu proteza prin intermediul propriilor semnale neuronale, primind informaţii despre presiune şi poziţie de la membru şi direcţionând mişcarea sa ulterioară. Aceasta s-a dovedit a fi o problemă mult mai dificilă decât părea pentru Wiener: după şaptezeci de ani, protezele care să încorporeze *feedback* neuronal sunt încă în stadii incipiente de dezvoltare. Conceptul lui Wiener a fost unul extraordinar, însă este dificil să creezi o interfaţă între semnalele neuronale şi dispozitivele electromecanice.

Mai important de atât, Wiener (alături de aproape toată lumea în anii 1950) a subapreciat masiv potenţialul tehnicii de calcul digitale. Aşa cum s-a mai observat, contribuţiile matematice ale lui Wiener au fost aduse pe palierul analizei semnalelor şi zgomotului, iar metodele sale analitice se aplică semnalelor cu variaţie continuă sau analoge. Deşi a participat la dezvoltarea tehnicii de calcul digitale în timpul războiului, nu a prevăzut niciodată explozia exponenţială a puterii de calcul determinată de introducerea şi miniaturizarea progresivă a circuitelor cu semiconductori. Însă nu este deloc vina lui Wiener: tranzistorul nu fusese inventat încă, şi tehnologia tuburilor vidate a computerelor digitale pe care o cunoştea el era grosolană, nefiabilă şi nu putea fi transferată la dispozitive din ce în ce mai mari. În anexa ediţiei din 1948 a *Ciberneticii*, anticipează computerele care vor juca şah şi preconizează că vor fi capabile să gândească cu două sau chiar trei mutări înainte. Poate ar fi fost surprins să afle că, după jumătate de secol, un computer]1 poate învinge pe campionul mondial uman la şah.

Supraestimarea tehnologiei şi riscurile existenţiale ale singularităţii

Când Wiener şi-a scris cartea, un exemplu semnificativ de supraestimare a tehnologiei era pe cale să se producă. Anii 1950 au văzut primele eforturi în dezvoltarea inteligenţei artificiale, din partea unor cercetători precum Herbert Simon, John McCarthy şi Marvin Minsky, care au început să programeze computere să realizeze sarcini simple şi au demarat construcţia de roboţi rudimentari. Succesul acestor eforturi iniţiale l-a inspirat pe Simon să declare că „maşinile vor fi capabile, în decurs de douăzeci de ani, să facă orice activitate pe care o poate face şi omul”. Astfel de predicţii s-au dovedit a fi spectaculos de greşite. Pe măsură ce au devenit mai puternice, computerele au devenit din ce în ce mai bune la jucat şah, pentru că puteau genera şi evalua sistematic o selecţie vastă de mutări posibile viitoare. Dar majoritatea predicţiilor despre inteligenţa artificială, de exemplu menajere robotice, s-au dovedit a fi iluzorii. Când Deep Blue

l-a învins pe Garry Kasparov la şah în 1997, cel mai puternic robot de curăţenie era Roomba, care se mişca şi aspira la întâmplare, scârţâind când rămânea blocat sub canapea.

Previziunile tehnologice sunt destul de periculoase, dat fiind faptul că tehnologiile progresează printr-o serie de îmbunătăţiri, încetinite de obstacole şi depăşite de inovaţie. Multe obstacole şi câteva inovaţii pot fi anticipate, dar şi mai multe nu pot. În activitatea mea cu experimentalişti care încearcă să construiască computere cuantice observ de obicei că anumite etape tehnologice de la care mă aştept să fie uşoare se dovedesc a fi imposibile, iar alte sarcini pe care mi le imaginez a fi imposibile se dovedesc a fi uşoare. Nu ştii până nu încerci.

În anii 1950, parţial inspirat de conversaţii cu Wiener, John von Neumann a introdus noţiunea de „singularitate tehnologică". Tehnologiile au tendinţa să se îmbunătăţească exponenţial, dublându-şi puterea sau sensibilitatea la un anumit interval de timp. (De exemplu, din 1950, tehnologiile computerelor şi-au dublat puterea la aproximativ doi ani, observaţie cunoscută ca Legea lui Moore.) Von Neumann a extrapolat rata exponenţială observată a îmbunătăţirilor tehnologice pentru a prezice că „progresul tehnologic va deveni incomprehensibil de rapid şi complicat", depăşind capacităţile omeneşti în viitorul nu prea îndepărtat. Într-adevăr, dacă extrapolăm creşterea puterii brute de calcul – exprimată în termeni de biţi şi bit-flips[6] – în viitor ca rata sa curentă, computerele-ar trebui să ajungă din urmă creierul omenesc în următoarele două până la patru decenii (în funcţie de modul de estimare a puterii de procesare a informaţiilor în creierele umane).

Eşecul predicţiilor iniţiale mult prea optimiste despre inteligenţa artificială a domolit discuţiile cu privire la singularitatea tehnologică timp de câteva decenii, dar, de la publicarea în anul 2005 a lucrării lui Ray Kurzweil *Singularitatea este aproape*[7], ideea progresului tehnologic care conduce la superinteligenţă a revenit în forţă. Unii adepţi ai ideii, printre care şi Kurzweil, privesc singularitatea drept o oportunitate: oamenii îşi pot contopi creierele cu superinteligenţă şi astfel pot trăi la nesfârşit. Alţii, de exemplu Stephen Hawking şi Elon Musk, au exprimat îngrijorare că această superinteligenţă s-ar putea dovedi malignă, privind-o ca pe cea mai mare ameninţare existentă la adresa civilizaţiei umane. Totuşi, alţii, printre care se numără şi câţiva participanţi la acest volum, consideră astfel de discuţii exagerate.

Munca de o viaţă a lui Wiener şi faptul că nu a prevăzut consecinţele acesteia sunt legate intim de ideea de singularitate tehnologică iminentă. Activitatea sa în domeniul neuroştiinţei şi sprijinul iniţial pe care l-a acordat lui McCulloch şi Pitts au umbrit metodele de învăţare profundă înspăimântător de eficiente din ziua de

azi. În ultimul deceniu şi în special în ultimii cinci ani, aceste tehnici de învăţare profundă au manifestat în sfârşit ceea ce Wiener numea *Gestalt* – de exemplu, capacitatea de a recunoaşte că un cerc este un cerc chiar şi atunci când este aplecat într-o parte şi arată ca o elipsă. Activitatea sa în domeniul comenzilor combinată cu activitatea din domeniul *feedbac*kului neuromuscular a fost semnificativă pentru dezvoltarea roboticii şi este inspiraţia pentru interfeţele dintre om şi maşină bazate pe neuroni. Scăpările sale în ceea ce priveşte predicţiile tehnologice însă sugerează că ar trebui să fim mai atenţi când discutăm despre noţiunea singularităţii tehnologice. Dificultăţile generale ale predicţiilor tehnologice şi problemele specifice dezvoltării unei superinteligenţe ar trebui să ne avertizeze cu privire la supraestimarea puterii şi eficienţei prelucrării informaţiilor.

Argumentele pentru scepticism în ceea ce priveşte singularitatea

Nicio creştere exponenţială nu durează la nesfârşit. O explozie atomică creşte exponenţial, dar numai cât timp are combustibil. Similar, avansurile exponenţiale din Legea lui Moore încep să atingă nişte limite impuse de fizica elementară. Frecvenţa procesoarelor computerelor a ajuns la un maximum de câţiva gigahertzi cu un deceniu şi jumătate în urmă, pur şi simplu pentru că cipurile au început să se topească. Miniaturizarea tranzistorilor a întâlnit deja obstacole cuantice şi mecanice din cauza fenomenului de *turmeling* şi a curenţilor de scurgere. În cele din urmă, multiplele îmbunătăţiri exponenţiale în ceea ce priveşte memoria şi procesarea guvernate de Legea lui Moore se vor opri. Însă, probabil, în câteva decenii, puterea brută de procesare a informaţiilor de care dispun computerele o va ajunge din urmă pe cea a creierului – cel puţin în măsura grosieră a numărului de biţi şi de *bit-fiips* pe secundă.

Creierele omeneşti sunt construite complex, prin procesul de selecţie naturală care a durat milioane de ani. Pe vremea lui Wiener, înţelegerea noastră în ceea ce priveşte arhitectura creierului era rudimentară şi simplistă. De atunci, instrumentaţia şi tehnicile imagistice din ce în ce mai sensibile au arătat că avem creiere cu o structură mult mai variată şi o funcţie mult mai complexă decât şi-a putut imagina Wiener. L-am întrebat recent pe Tomaso Poggio, unul dintre pionierii neuroştiinţei moderne, dacă se îngrijorează de faptul că funcţia creierului uman va fi emulată curând de computere, dată fiind puterea lor crescândă de procesare a informaţiei. „Nicio şansă”, mi-a răspuns.

Progresele recente ale învăţării profunde şi ale tehnicii de calcul neuromorfe sunt foarte bune la reproducerea unui anumit aspect al inteligenţei omeneşti concentrat pe funcţionarea cortexului, acolo

unde tiparele sunt procesate şi recunoscute. Aceste progrese au permis unui computer să învingă nu numai campionul mondial la şah, cât şi la Go, o realizare impresionantă, dar mai e drum lung până la robotul computerizat care să-ţi facă curat în cameră. (De fapt, roboţii care să se apropie de vasta capacitate umană a gamei de mişcări flexibile sunt încă un vis îndepărtat – căutaţi „roboţi căzând”[8] pe internet. Roboţii sunt performanţi când trebuie să facă suduri de precizie pe liniile de asamblare, dar încă nu se pot lega la şireturi.)

Puterea brută de procesare a informaţiilor nu presupune o putere sofisticată de procesare a informaţiilor. Deşi puterea computerelor a avansat exponenţial, programele cu care funcţionează computerele adesea nu au avansat deloc. Unul dintre principalele răspunsuri ale companiilor de software la puterea de procesare sporită este să adauge caracteristici „utile”, care fac adesea softul mai greu de utilizat. Microsoft Word şi-a atins punctul culminant în 1995 şi, de atunci, se scufundă încet sub greutatea caracteristicilor adăugate. Când Legea lui Moore va începe să încetinească, dezvoltatorii de software vor fi nevoiţi să facă alegeri dificile între eficientă, viteză şi funcţionalitate.

O temere majoră a singularitariştilor[9] este că, pe măsură ce computerele se vor implica din ce în ce mai mult în proiectarea propriilor softuri, îşi vor încărca rapid instrucţiuni iniţiale care le vor permite să atingă abilităţi de calcul supraomeneşti. Însă dovezile pe care ni le pune la dispoziţie învăţarea automată arată altfel. Pe măsură ce maşinile devin mai puternice şi capabile de învăţare, învaţă din ce în ce mai mult ca oamenii – din exemple multiple, adesea sub supravegherea instructorilor umani sau electronici. Educaţia este la fel de dificilă şi lentă pentru computere cum este şi pentru adolescenţi. În consecinţă, sistemele bazate pe învăţare profundă devin mai omeneşti şi nu mai puţin omeneşti. Competenţele pe care le aduc învăţării nu sunt „mai bune decât”, ci „complementare” învăţării umane: sistemele computerizate de învăţare pot identifica tipare pe care oamenii nu le pot identifica – şi invers. Cei mai buni jucători de şah din lume nu sunt nici computere şi nici oameni, ci oameni care lucrează împreună cu computerele. Cyberspaţiul este într-adevăr populat de programe dăunătoare, dar acestea iau preeminent forma de *malware* – viruşi notabili pentru lipsa lor malignă de raţiune, şi nu pentru superinteligenţa lor.

Încotro, Wiener?

Wiener a observat că progresul tehnologic exponenţial este un fenomen relativ modern şi nu este pe de-a-ntregul bun. Privea armele atomice şi dezvoltarea rachetelor cu focoase nucleare ca pe o reţetă pentru sinuciderea rasei umane. A comparat exploatarea impetuoasă a resurselor planetei cu petrecerea cu ceai a Pălărierului Nebun din *Alice*

în Ţara Minunilor. Pustiind mediul local dintr-un anumit loc, progresăm pur şi simplu mutându-ne în alt loc şi pustiindu-l şi pe acela. Optimismul lui Wiener cu privire la dezvoltarea computerelor şi sistemelor neuromecanice a fost temperat de pesimismul său cu privire la exploatarea acestora de către guvernele asupritoare, cum ar fi cazul Uniunii Sovietice, şi la tendinţa democraţiilor, cum ar fi cea din Statele Unite ale Americii, să devină şi ele mai autoritare când se confruntă cu ameninţarea autoritarismului.

Ce ar crede Wiener despre folosirea omului de către om din prezent? Ar fi uluit de puterea computerelor şi a internetului. S-ar bucura că reţelele neuronale incipiente, în care a avut un rol important, au dat naştere unor sisteme puternice de învăţare profundă care manifestă capacitatea perceptivă pe care o aştepta de la ele, deşi, probabil, nu ar fi impresionat de faptul că unul dintre cele mai remarcabile exemple ale acestui *Gestalt* computerizat este capacitatea de a recunoaşte fotografii cu pisici pe World Wide Web. În loc să privească inteligenţa maşinilor ca pe o ameninţare, cred că ar privi-o ca pe un fenomen autonom, diferit deşi coevoluând cu propria noastră inteligenţă umană.

Nesurprins de încălzirea globală – petrecerea cu ceai a Pălărierului Nebun din era noastră –, Wiener ar aplauda îmbunătăţirea exponenţială a tehnologiilor de energie alternativă, aplicându-şi cunoştinţele cibernetice în dezvoltarea setului complex de cicluri de *feedback* necesar pentru încorporarea acestor tehnologii în reţeaua electrică inteligentă ce urmează să apară. Cu toate acestea, recunoscând că soluţia problemei încălzirii globale este cel puţin în aceeaşi măsură politică, pe cât este tehnologică, fără îndoială că ar fi pesimist gândindu-se la şansele noastre de a rezolva această problemă care ne ameninţă civilizaţia în timp. Wiener detesta afaceriştii – în special pe cei din lumea politică –, dar recunoştea că aceştia vor fi întotdeauna cu noi.

Este uşor să uităm cât de înspăimântătoare era lumea lui Wiener. Statele Unite ale Americii şi Uniunea Sovietică erau în plină cursă a înarmărilor, construind bombe cu hidrogen montate pe focoase nucleare transportate de rachete balistice intercontinentale ghidate de sisteme de navigare la care contribuise Wiener însuşi – spre disperarea sa. Aveam patru ani când a murit Wiener. În 1964, grădiniţa mea se pregătea pentru un atac nuclear punându-ne să ne ascundem sub bănci. Dată fiind folosirea omului de către om din vremea sa, dacă ar putea vedea starea noastră curentă, primul răspuns al lui Wiener ar fi să răsufle uşurat că suntem încă în viaţă.

Capitolul 2

JUDEA PEARL

Judea Pearl este profesor de informatică şi administrator al Laboratorului de Sisteme Cognitive de la UCLA. Cea mai recentă carte a sa, scrisă împreună cu Dana Mackenzie, este The Book of Why: The New Science of Carne and Efect.

În anii 1980, Judea Pearl a introdus o abordare nouă a inteligenţei artificiale denumită reţelele bayesiene. Acest model bazat pe probabilităţi de raţionare automată a permis maşinilor să funcţioneze – într-o lume complexă şi nesigură – ca „motoare de dovezi", revizuindu-şi continuu concluziile în lumina noilor dovezi.

În decurs de câţiva ani, reţelele bayesiene ale lui Judea au întrecut complet abordările anterioare ale inteligenţei artificiale bazate pe reguli. Apariţia învăţării profunde – în care computerele, practic, se învaţă singure să fie mai inteligente prin observarea multor date – l-a lăsat puţin sceptic, pentru că această metodă nu este transparentă.

Recunoscând realizările impresionante în domeniul învăţării profunde din partea unor colegi precum Michael I. Jordan şi Geoffrey Hinton, se simte inconfortabil cu acest tip de opacitate. Şi-a propus să înţeleagă limitările teoretice ale sistemelor de învăţare profundă şi evidenţiază faptul că există bariere elementare care-i vor împiedica să obţină un tip de inteligenţă omenească, indiferent ce vor face. Folosind beneficiile de calcul ale reţelelor bayesiene, Judea şi-a dat seama că o combinaţie de simple modele grafice şi date poate fi utilizată şi pentru a reprezenta şi deduce relaţiile cauză-efect. Importanţa descoperirii sale trece cu mult dincolo de rădăcinile sale în inteligenţa artificială. Ultima sa carte explică gândirea cauzală pentru publicul larg; am putea spune că este un prim nivel despre cum gândim, chiar şi pentru oameni.

Perspectiva principială, matematică a lui Judea asupra cauzalităţii, este o contribuţie profundă pentru domeniul ideilor. A adus deja beneficii aproape fiecărui domeniu al cercetării, în special în ştiinţele sociale şi medicale, care se bazează pe un volum mare de date.

În calitate de fost fizician, eram foarte interesat de cibernetică. Deşi nu utiliza întreaga putere a maşinilor Turing, era foarte transparentă, probabil deoarece era fondată pe teoria clasică a controlului şi teoria informaţiei. Pierdem acum din această transparenţă, odată cu stilul de învăţare profundă al învăţării

automate. Fundamental, este un exerciţiu de ajustare a curbei ce reglează ponderile nivelurilor intermediare dintr-un lanţ lung de intrare-ieşire.

Întâlnesc mulţi utilizatori care spun că „funcţionează bine şi nu ştim de cca. Când este dezlănţuită asupra seturilor mari de date, învăţarea profundă are propria sa dinamică, îşi face propriile reparaţii şi optimizări şi îţi oferă rezultatele corecte de cele mai multe ori. Dar, atunci când nu reuşeşte să facă asta, nu ai nicio idee despre ce nu a mers şi cum poţi să rezolvi problema, îndeosebi nu ştii dacă eroarea este în program, în metodă sau cauzată de schimbarea stării mediului. Ar trebui să urmărim un alt fel de transparenţă.

Unii argumentează că transparenţa nu este neapărat necesară. Nu înţelegem arhitectura neuronală a creierului uman, însă funcţionează bine, aşa că uităm de înţelegerea noastră superficială şi folosim ajutoarele umane pentru avantaje uriaşe. În anumite feluri, argumentează ei, de ce să nu dezlănţuim sisteme de învăţare profundă şi să creăm inteligenţă fără să înţelegem cum funcţionează? Sunt de acord cu acest argument într-o anumită măsură. Personal, nu-mi place opacitatea, aşa că nu-mi voi irosi timpul cu învăţarea profundă, dar ştiu că-şi are locul în compoziţia inteligenţei. Ştiu că sistemele netransparente pot face minuni şi creierul nostru este dovada vie a acestor minuni.

Însă acest argument are limitările sale. Motivul pentru care putem să iertăm înţelegerea noastră superficială în ceea ce priveşte funcţionarea creierului uman este că creierele noastre funcţionează la fel, şi acest lucru ne permite să comunicăm cu alţi oameni, să învăţăm de la ei, să-i instruim şi să-i motivăm în limba noastră maternă. Dacă roboţii noştri vor fi toţi la fel de opaci ca Alpha Go, nu vom putea susţine conversaţii profunde cu ei şi ar fi trist. Ar trebui să-i reinstruim de câte ori am face o mică schimbare sarcinii sau mediului de funcţionare.

Aşadar, decât să experimentez cu maşinile de învăţare opace, încerc să înţeleg limitările lor teoretice şi să examinez cum pot fi depăşite aceste limitări. Fac acest lucru în contextul sarcinilor de raţiune cauzală, care guvernează o mare parte din perspectiva oamenilor de ştiinţă asupra lumii, şi, în acelaşi timp, sunt bogate în intuiţie şi în exemple, pentru a putea monitoriza progresul analizei noastre. În acest context, am descoperit că există câteva bariere elementare şi, dacă nu le depăşim cumva, nu ne vom apropia măcar de un soi de inteligenţă umană, indiferent ce facem. După mine, cartografierea acestor bariere nu este mai puţin importantă decât să ne lovim de ele.

Sistemele curente de învăţare automată funcţionează aproape exclusiv într-un mod statistic sau cu model orb, lucru analog în multe

feluri încercării de a implementa o funcţie într-un *cloud* de puncte de date. Astfel de sisteme nu îşi pot pune întrebarea „Ce-ar fi dacă?”, neputând servi aşadar drept bază pentru o inteligenţă artificială puternică – şi anume, o inteligenţă artificială care emulează raţiunea şi competenţa umane. Pentru a ajunge la nivelul inteligenţei umane, maşinile de învăţare au nevoie de ghidarea unei scheme a realităţii, unui model – similar unei hărţi care ne ghidează atunci când conducem printr-un oraş pe care nu-l cunoaştem.

Pentru a fi mai clar, maşinile curente de învăţare îşi îmbunătăţesc performanţa optimizând parametri pentru un flux de intrări senzoriale primite de la mediu. Este un proces lent, analog procesului de selecţie naturală care stă la baza evoluţiei darwiniste. Explică, astfel, cum specii precum vulturii şi şerpii au dezvoltat sisteme de văz excelente de-a lungul a milioane de ani, însă nu poate explica procesul supraevoluţionar care permite oamenilor să producă ochelari şi telescoape în decurs de mai puţin de o mie de ani. Ce au avut oamenii şi nu au avut alte specii a fost o reprezentare mentală a mediului lor – o reprezentare pe care au putut să o manipuleze după propria voinţă pentru a-şi imagina medii ipotetice alternative pentru planificare şi învăţare.

Istorici ai lui homo sapiens, cum ar fi Yuval Noah Harari şi Steven Mithen, sunt de acord că ingredientul decisiv care a conferit strămoşilor noştri capacitatea de a obţine dominaţia asupra întregii planete, cu aproximativ 40.000 de ani în urmă, a fost capacitatea lor de a crea şi memora o reprezentare mentală a mediului înconjurător, de a interoga acea reprezentare, de a o distorsiona prin acte mentale ale imaginaţiei şi de a răspunde în cele din urmă întrebărilor care încep cu „Dar dacă?” Exemple sunt întrebările intervenţioniste (Dacă fac aşa şi aşa?”) şi întrebările retrospective şi contrafactuale (Dacă aş fi făcut altceva?”). Nicio maşină de învăţare în funcţiune astăzi nu poate răspunde la aceste întrebări. Mai mult chiar, majoritatea maşinilor de învăţare nu posedă o reprezentare din care să derive răspunsurile la astfel de întrebări.

Cu privire la raţiunea cauzală, constatăm că poţi face foarte puţine lucruri cu orice fel de ajustare a curbei cu model orb sau orice inferenţă statistică, indiferent cât este de sofisticat procesul de ajustare. De asemenea, am găsit un cadru teoretic pentru organizarea acestor limitări, care formează o ierarhie.

La primul nivel, ai gândirea statistică, ce-ţi poate spune doar cum percepţia unui eveniment îţi poate schimba perspectiva asupra altuia. De exemplu, ce-ţi poate spune un simptom despre o boală?

Apoi ai un al doilea nivel, care-l implică pe primul, dar nu şi invers. Aici este vorba despre acţiuni. „Ce s-ar întâmpla dacă am creşte preţurile?” „Dacă mă faci să râd?” Acest al doilea nivel al

ierarhiei necesită informaţii despre intervenţiile care nu sunt disponibile în primul. Aceste informaţii pot fi codate într-un model grafic, care ne spune doar ce variabilă corespunde alteia.

Al treilea nivel al ierarhiei este cel contrafactual. Acesta este limbajul utilizat de oamenii de ştiinţă. „Dacă obiectul ar fi de două ori mai greu?" „Dacă aş face lucrurile altfel?" „Durerea de cap mi-a trecut de la aspirină sau de la somn?" Enunţurile contrafactuale sunt pe ultimul nivel în sensul că nu pot fi derivate nici dacă am putea prezice efectele tuturor acţiunilor. Au nevoie de un ingredient suplimentar, sub forma ecuaţiilor, pentru a ne spune cum răspund variabilele la modificările altor variabile.

Una dintre cele mai mari realizări ale cercetării inferenţei cauzale a fost algoritmizarea intervenţiilor şi enunţurilor contrafactuale, ultimele două niveluri ale ierarhiei. Cu alte cuvinte, după ce codăm cunoştinţele noastre ştiinţifice într-un model (care poate fi calitativ), există algoritmi care examinează modelul şi stabilesc dacă o anumită interogaţie, fie ea despre intervenţie sau despre enunţul contrafactual, poate fi estimată din datele disponibile – şi, dacă da, cum. Această capacitate a transfigurat modul în care oamenii de ştiinţă se ocupă de ştiinţă, în special în cele în care informaţiile abundă, cum sunt sociologia şi epidemiologia, pentru care modelele cauzale au devenit o a doua limbă. Aceste discipline văd transformările lor lingvistice ca pe Revoluţia Cauzală. Cum spune cercetătorul de ştiinţe sociale de la Harvard, Gary King, „s-au învăţat mai multe despre inferenţa cauzală în ultimele decenii decât s-a învăţat în toată istoria înregistrată până atunci".

Contemplând succesul învăţării automate şi încercând să-l extrapolez la viitorul inteligenţei artificiale, mă întreb: suntem conştienţi de limitările elementare care au fost descoperite în arena inferenţei cauzale? Suntem pregătiţi să facem faţă impedimentelor teoretice care ne împiedică să trecem de la un nivel al ierarhiei la altul?

Eu privesc învăţarea automată ca pe un instrument care ne conduce de la date la probabilităţi. Dar apoi trebuie să mai facem încă doi paşi pentru a trece de la probabilităţi la înţelegerea reală – doi paşi mari. Unul este să preconizăm efectul acţiunilor, iar al doilea este imaginaţia contrafactuală. Nu putem pretinde că înţelegem realitatea decât dacă facem ultimii doi paşi.

În pătrunzătoarea sa carte *Foresight and Understanding* (1961), filosoful Stephen Toulmin a identificat contrastul dintre transparenţă şi opacitate drept cheia înţelegerii rivalităţii antice dintre ştiinţele greacă şi babiloniană. După părerea lui Toulmin, astronomii babilonieni erau maeştri ai predicţiilor de tip cutie neagră, întrecându-i lejer pe rivalii lor greci în ceea ce priveşte acurateţea şi concordanţa

observaţiilor celeste. Însă ştiinţa a favorizat strategia creativ-speculativă a astronomilor greci, care era întreţesută puternic cu imagistica metaforică: tuburi circulare pline de foc, orificii mici prin care focul celest era vizibil sub formă de stea şi pământul emisferic purtat pe spatele unei broaşte-ţestoase uriaşe. Această strategie de modelare excentrică, şi nu extrapolarea babiloniană a fost cea care i-a dat imboldul lui Eratostene (276 – 194 î.Hr.) să realizeze unul dintre cele mai creative experimente din lumea antică şi să calculeze circumferinţa Pământului. Un astfel de experiment nu i-ar fi trecut niciodată prin minte unui arhivator de date babilonian.

Abordările cu model orb impun limitări intrinseci sarcinilor cognitive pe care inteligenţa artificială puternică le poate îndeplini. Concluzia mea generală este că inteligenţa artificială de nivel uman nu poate lua naştere doar din maşini de învăţare cu model orb; necesită colaborarea simbiotică a datelor şi modelelor.

Ştiinţa datelor este o ştiinţă doar în măsura în care facilitează interpretarea datelor – o problemă cu două faţete, conectarea datelor la realitate. Datele în sine nu sunt o ştiinţă, indiferent cât de „mari" ajung şi de iscusinţa cu care sunt manipulate. Sistemele opace de învăţare ne pot duce până în Babilon, dar nu până în Atena.

Capitolul 3
Scopul scris în maşină

Stuart Russell

Stuart Russell este profesor de informatică şi profesor de inginerie Smith-Zadeh la Universitatea Berkeley. Este coautorul (alături de Peter Norvig) al cărţii Artificial Intelligence: A Modern Approach.

Informaticianul Stuart Russell alături de Elon Musk, Stephen Hawking, Max Tegmark şi mulţi alţii au insistat să se acorde atenţie potenţialelor pericole implicate de crearea unei inteligenţe de nivel supraomenesc (sau chiar omenesc) – o inteligenţă artificială generală, ale cărei scopuri programate nu se vor alinia neapărat cu ale noastre.

Cărţile sale anterioare au fost pe tema înţelegerii noţiunii de „caracter optim legat" ca definiţie formală a inteligenţei pe care se poate lucra. A dezvoltat tehnica metaraţionării raţionale, „care este, în termeni largi, de a face calculele aşteptate pentru îmbunătăţirea calităţii deciziei tale finale cât se poate de repede". De asemenea, a lucrat la unificarea teoriei probabilităţii şi logicii de gradul întâi – ceea ce a condus la crearea unui sistem de

monitorizare nou şi mult mai eficient pentru Tratatul privind interzicerea totală a testelor nucleare – şi la problema procesului de luare a deciziilor pe perioade lungi de timp (prezentările sale pe această ultimă temă se intitulează, de obicei „Viaţa: joacă şi câştigă în 20 de trilioane de mutări").

Este foarte îngrijorat de dezvoltarea continuă a armelor autonome, cum ar fi microdronele letale, care pot fi transformate în arme de distrugere în masă. A redactat o scrisoare de la patruzeci de cercetători de top din domeniul inteligenţei artificiale şi a trimis-o preşedintelui Obama, determinând astfel o serie de şedinţe de securitate naţională la nivel înalt.

Activitatea sa din prezent se concentrează pe crearea a ceea ce el numeşte inteligenţă artificială „vădit benefică". Doreşte să creeze siguranţa inteligenţei artificiale „impregnând sistemele cu nesiguranţă explicită" cu privire la obiectivele programatorilor lor umani, o abordare care ar presupune o reorganizare radicală transparentă a cercetării inteligenţei artificiale din prezent.

Stuart este, de asemenea, şi pe radarul oricărei persoane care a făcut un curs universitar de informatică în ultimii douăzeci şi ceva de ani. Este coautorul celui mai cunoscut manual pe tema inteligenţei artificiale, cu un număr estimat de peste 5 milioane de cititori în limba engleză.

Printre multele probleme semnalate în cartea lui Norbert Wiener *Folosirea omului de către om* (1950) care sunt relevante în prezent, cea mai importantă pentru cercetătorul inteligenţei artificiale este posibilitatea ca umanitatea să cedeze maşinilor controlul asupra destinului său.

Wiener a considerat maşinile din viitorul apropiat mult prea limitate pentru a exercita un control global, imaginându-şi, în schimb, că maşinile şi sistemele de control asemănătoare maşinilor vor fi controlate de elitele omeneşti pentru a reduce marea masă a umanităţii la starea de „rotiţe, manete şi tije". Privind şi mai mult în viitor, a evidenţiat dificultatea de a specifica absolut corect obiectivele pentru maşinile cu capacităţi mari, observând

„Adevăruri simple şi evidente ale vieţii, cum ar fi djinnul găsit într-o sticlă, pe care ai face bine să-l laşi acolo; al pescarului care-şi doreşte prea multe binefaceri de la Dumnezeu pentru soţia sa şi care ajunge fix acolo de unde a plecat, al faptului că, dacă ţi se acordă trei dorinţe, trebuie să fii foarte atent ce-ţi doreşti".

Pericolele sunt clare:

„Vai de noi dacă lăsăm «maşina» să ne decidă comportamentul înainte de a examina legile acţiunilor sale şi de a

şti bine dacă acest comportament al maşinii va fi ghidat de principii acceptabile pentru noi! Pe de altă parte, maşina, asemenea djinnului, care poate învăţa şi lua decizii în baza a ceea ce a învăţat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi".

Zece ani mai târziu, după ce a văzut cum programul de jucat table al lui Arthur Samuel învaţă să joace table mult mai bine decât creatorul său, Wiener a publicat articolul „Câteva consecinţe morale şi tehnice ale automatizării" în revista *Science*. În această lucrare, mesajul este şi mai clar:

„Dacă, pentru a ne atinge scopurile, folosim un agent mecanic cu a cărui funcţionare nu putem interfera în mod eficient… ar fi bine să fim foarte siguri că scopul pe care îl programăm în maşini este scopul pe care-l dorim cu adevărat".

Din punctul meu de vedere, aceasta este sursa riscului existenţial al inteligenţelor artificiale superinteligente citat în ultimii ani de observatori precum Elon Musk, Bill Gates, Stephen Hawking şi Nick Bostrom.

Programarea scopurilor în maşini

Obiectivul cercetării inteligenţei artificiale a fost înţelegerea principiilor care stau la baza comportamentului inteligent şi integrarea acelor principii în maşini care pot manifesta un astfel de comportament. În anii 1960 şi 1970, noţiunea prevalentă de inteligenţă era capacitatea de gândire logică, ce includea abilitatea de a deriva planuri de acţiune ce puteau atinge garantat un obiectiv specificat. Recent, s-a creat un consens în jurul ideii de agent raţional care percepe şi acţionează în conformitate pentru a-şi maximiza utilitatea prevăzută. Subdomenii, cum ar fi planificarea logică, robotica şi înţelegerea limbajului natural, sunt cazuri speciale de paradigme generale. Inteligenţa artificială a încorporat teoria probabilităţii pentru a face faţă nesiguranţei, teoria utilităţii pentru a defini obiectivele şi învăţarea statistică pentru a permite maşinilor să se adapteze la circumstanţe noi. Aceste dezvoltări au creat legături puternice cu alte discipline care se bazează pe concepte similare, printre care teoria controlului, economia, cercetarea operaţiilor şi statistica.

Atât din perspectiva de planificare logică, precum şi din cea de agent raţional asupra inteligenţei artificiale, obiectivul maşinii – fie acesta sub formă de scop, funcţie utilitară sau funcţie de recompensare (ca în învăţarea cu întărire) – este specificat din exterior. În cuvintele lui Wiener, acesta este „scopul programat în maşină". Într-adevăr, este

una dintre regulile de bază ale domeniului ca sistemele de inteligenţă artificială să servească scopuri generale – adică să fie capabilă să accepte un scop ca intrare şi apoi să-l atingă –, şi nu scopuri speciale, cu obiectivul stabilit implicit în designul lor. De exemplu, un autoturism cu pilot automat ar trebui să accepte o destinaţie ca intrare, şi nu să aibă o destinaţie fixă. Însă unele aspecte ale „scopului de şofat" ale autoturismului sunt fixe, cum ar fi faptul că nu trebuie să lovească pietonii. Acesta este construit direct în algoritmii de şofat ai autoturismului, nefiind explicit: niciun autoturism cu pilot automat care există în zilele noastre nu „ştie" că pietonii preferă să nu fie călcaţi.

Programarea unui scop într-o maşină pentru a-i optimiza comportamentul conform unor algoritmi clar definiţi pare să fie o abordare admirabilă pentru a ne asigura că, aşadar, „comportamentul maşinii se va desfăşură conform unor principii acceptabile pentru noi!" Dar, aşa cum avertizează Wiener, trebuie să introducem scopul potrivit. Am putea denumi această problemă problema regelui Midas: Midas a primit exact ce a cerut – şi anume ca tot ce atinge să se transforme în aur –, dar a descoperit prea târziu dezavantajele de a bea aur lichid şi de a mânca aur solid. Termenul tehnic pentru programarea scopului potrivit este alinierea valorii. Când eşuează, putem insufla accidental maşinilor obiective contrare cu ale noastre. Dacă sistemul de inteligenţă artificială primeşte sarcina de a găsi un tratament pentru cancer cât se poate de repede, acesta ar putea alege să folosească întreaga populaţie umană drept cobai pentru experimentele sale. Dacă i se cere să amelioreze aciditatea din oceane, ar putea folosi tot oxigenul din atmosferă ca efect secundar. Aceasta este o caracteristică obişnuită a sistemelor care folosesc optimizarea: variabilele care nu sunt incluse în obiectiv pot fi setate la valori extreme pentru a ajuta la optimizarea obiectivului respectiv.

Din nefericire, nici inteligenţa artificială şi nici alte discipline (economia, statistica, teoria controlului, cercetarea operaţiilor) structurate în jurul optimizării obiectivelor nu au multe de spus despre metoda de identificare a scopurilor „pe care ni le dorim cu adevărat". În schimb, acestea presupun că obiectivele sunt pur şi simplu implantate în maşină. Cercetarea inteligenţei artificiale, în forma sa actuală, studiază abilitatea de a atinge obiective, şi nu proiectarea obiectivelor respective.

Steve Omohundro a arătat încă o dificultate, observând că entităţile inteligente trebuie să acţioneze pentru a-şi prezerva propria existenţă. Această tendinţă nu are nimic de-a face cu instinctul de autoconservare sau altă noţiune biologică; pur şi simplu, o entitate nu-şi poate atinge obiectivul dacă este moartă. Conform argumentului lui Omohundro, o maşină superinteligentă care are un comutator de

oprire – pe care unii, inclusiv Alan Turing într-o prelegere susţinută în 1951 la BBC Radio 3, au văzut-o ca pe potenţiala noastră salvare – va lua măsuri pentru a dezactiva acest comutator într-o anumită măsură.[10] Astfel putem face faţă ideii de maşini superinteligente – acţiunile lor fiind prin definiţie impredictibile pentru noi, iar obiectivele lor imperfect specificate intrând în conflict cu ale noastre –, ale căror motivaţii de a-şi prezerva existenţa pentru a atinge acele obiective ar putea fi de neînvins.

1001 motive de a nu îi atenţi

Au fost emise obiecţii la aceste argumente, în primul rând de către cercetătorii din comunitatea inteligenţei artificiale. Obiecţiile reflectă o reacţie defensivă naturală, cuplată probabil cu lipsa imaginaţiei despre ce ar putea face o maşină superinteligentă. Niciuna nu pare să rămână în picioare la o examinare mai atentă. Iată câteva dintre cele mai comune:

- *Nu-ţi face probleme, o putem dezactiva pur şi simplu. Acesta este adesea primul lucru care apare în mintea unui diletant când ia în considerare riscurile inteligenţei artificiale superinteligente – ca şi cum o entitate superinteligentă nu s-ar gândi niciodată la asta. Aceasta se aseamănă mai degrabă cu a spune că riscul de a pierde în faţa lui Deep Blue sau Alpha Go este neglijabil – tot ce trebuie să facem este să facem mutările potrivite.*
- *Inteligenţa artificială de nivel omenesc sau supraomenesc este imposibila" Aceasta este o afirmaţie neobişnuită pentru cercetătorii din domeniul inteligenţei artificiale, dat fiind faptul că, de la Turing încoace, s-au apărat de asemenea acuzaţii venite de la filosofi şi matematicieni. Afirmaţia, care nu are la bază nicio dovadă, pare să accepte că, dacă inteligenţa artificială superinteligentă ar fi posibilă, ar fi un risc semnificativ. Este ca şi cum un şofer de autobuz, având ca pasageri toată umanitatea, ar spune: „Da, conduc către o[11] prăpastie – de fapt, o să apăs pedala până la blană! Dar ai încredere în mine, o să rămânem fără combustibil înainte de a ajunge!" Afirmaţia reprezintă un pariu nesăbuit contra ingeniozităţii umane. Am mai făcut astfel de pariuri şi am pierdut. În data de 11 septembrie 1933, renumitul fizician Ernest Rutherford a afirmat cu încredere deplină: „Oricine se aşteaptă să descoperim o sursă de energie în transformarea acestor atomi*

vorbeşte prostii". În data de 12 septembrie 1933, Leo Szilard a inventat reacţia nucleară în lanţ indusă de neutroni. La câţiva ani, a demonstrat o astfel de reacţie în laboratorul său de la Universitatea din Columbia. După cum îşi amintea într-un memoriu: „Am oprit totul şi am mers acasă. În noaptea aceea, în mintea mea mai încăpea foarte puţină îndoială că lumea se îndrepta către suferinţă".

- *Este prea devreme să nefacem probleme. Momentul potrivit să ne îngrijorăm de o potenţială problemă gravă a umanităţii nu depinde numai de momentul în care va apărea problema, ci şi de cât timp este necesar pentru a dezvolta şi implementa o soluţie care evită riscurile. De exemplu, dacă ar fi să detectăm un asteroid uriaş care urmează să lovească Pământul în 2067, am spune oare: „Este prea devreme să ne îngrijorăm"? Şi, dacă analizăm riscurile globale catastrofale ale încălzirii globale care urmează să aibă loc la sfârşitul acestui secol, este oare prea devreme să luăm măsuri pentru a le preveni? Dimpotrivă, este posibil să fie prea târziu. Perioada de timp relevantă pentru inteligenţa artificială de nivel omenesc este mai puţin predictibilă, dar, ca fisiunea nucleară, poate sosi mult mai devreme decât ne aşteptăm. O variaţie a acestui argument este declaraţia lu' Andrew Ng că ar fi „ea şi cum ne-am îngrijora de suprapopularea planetei Marte". Aceasta trimite la o analogie convenabilă: nu doar că riscul este uşor de gestionat şi departe în viitor, dar este şi extrem de improbabil să încercăm să mutăm miliarde de oameni pe Marte. Analogia este falsă, însă. Deja dedicăm resurse tehnice şi ştiinţifice considerabile creării sistemelor de inteligenţă artificială din ce în ce mai performante. O analogie mai potrivită ar fi să plănuim să mutăm rasa umană pe Marte fără a lua în considerare ce vom mânca, ce vom bea şi ce vom respira după ce ajungem.*
- *Inteligenţa artificială de nivel omenesc nu este iminentă, în orice caz. Raportul AI100, de exemplu, ne asigură: „Contrar preconizărilor fantastice pentru inteligenţa artificială în presa populară, Comitetul de Studiu nu a găsit nicio cauză de îngrijorare care să arate că inteligenţa artificială este o ameninţare iminentă pentru omenire". Acest argument pur şi simplu greşeşte motivele de îngrijorare, care nu sunt predicate pe iminenţă. În cartea sa din 2014, Superintelligence:*

Paths, Dangers, Strategies, Nick Bostrom, de exemplu, scrie: „Nu face parte din argumentaţia acestei cărţi că suntem pe cale să facem o descoperire revoluţionară în inteligenţa artificială sau că putem prezice cu precizie când ar putea apărea o astfel de dezvoltare”.

- *Eşti doar un lucidit. Este o definiţie ciudată a luciditului care-i include pe Turing, Wiener, Minsky, Musk şi Gates, care sunt printre cei mai prolifici participanţi la progresul tehnologic din secolele al XX-lea şi al XXI-lea[12]. În plus, epitetul reprezintă o neînţelegere completă a naturii îngrijorărilor emise şi a scopului pentru care au fost emise. Este ca şi cum cineva ar acuza inginerii nucleari de luddism dacă ar evidenţia nevoia de a controla reacţia de fisiune. Unii contestatari folosesc şi termenul „anti-AI”, care este ca şi cum ai spune despre inginerii nucleari că sunt „antifizică”. Scopul înţelegerii şi împiedicării riscurilor inteligenţei artificiale este să ne asigurăm că ne putem bucura de beneficii. Bostrom, de exemplu, scrie că succesul controlării inteligenţei artificiale va avea ca efect „o traiectorie civilizaţională care va conduce la o folosire compasională şi jubilantă a dotărilor cosmice ale umanităţii” – o predicţie deloc pesimistă.*
- *Orice maşină care este suficient de inteligentă să cauzeze probleme va fi suficient de inteligentă să aibă obiective corespunzătoare şi altruiste. (Adesea, argumentul adaugă premisa că oamenii cu inteligenţă superioară au tendinţa de a avea obiective mai altruiste, o perspectivă care poate fi legată mai mult de şinele celor care aduc argumentul.) Acest argument este legat de problema lui Hume privind discrepanţa între ceea ce este şi ceea ce ar trebui să fie[13] şi eroarea naturalistă a lui G.E. Moore, care sugerează că maşina, ca urmare a inteligenţei sale, va percepe cumva ce este bine, dată fiind experienţa ei în lume. Acest lucru este neplauzibil; de exemplu, un om nu poate deduce obiectivul de şah mat privind tabla de şah sau piesele de şah; aceeaşi tablă de şah şi aceleaşi piese pot fi utilizate pentru antişah sau multe alte jocuri care nu au fost inventate încă. Cu alte cuvinte: în timp ce Bostrom îşi imaginează oamenii duşi la extincţie de un robot imaginar care transformă planeta într-o mare de agrafe pentru hârtie, noi oamenii vedem acest rezultat ca tragic, în timp ce bacteria thiobacillus ferrooxidans care consumă fier este*

încântată. Cine poate spune că bacteria greşeşte? Faptul că maşina a primit un obiectiv clar din partea oamenilor nu înseamnă că va recunoaşte automat importanţa oamenilor ca lucruri care nu fac parte din obiectiv. Maximizarea obiectivului poate cauza probleme oamenilor, dar, prin definiţie, maşina nu va recunoaşte acele probleme ca problematice.

- *Inteligenţa este multidimensionala, deci nmai inteligentă decât a oamenilor" este un conceptfără sens. Este o idee principală a psihologiei moderne că IQ-ul nu reprezintă întregul spectru de competenţe cognitive pe care le posedă oamenii în diverse grade. IQ-ul este într-adevăr o măsurătoare lipsită de precizie a inteligenţei umane, dar nu are niciun sens pentru sistemele moderne de inteligenţă artificială, deoarece capacităţile lor din diverse domenii nu sunt corelate. Cum putem compara IQrul motorului de căutare Google, care nu poate juca şah, cu cel al lui Deep Blue, care nu poate răspunde cererilor de căutare?*
- *Niciun motiv nu sprijină argumentul conform căruia, dacă inteligenţa este multilaterală, putem ignora riscul maşinilor superinteligente. Dacă „mai inteligent decât oamenii" este un concept fără sens, atunci la fel este şi „mai inteligent decât gorilele" şi, prin urmare, gorilele nu au niciun motiv de îngrijorare în ceea ce-i priveşte pe oameni. În mod clar argumentul nu este valabil. Nu doar că este posibil pentru o entitate să fie mai capabilă decât alta în toate dimensiunile relevante ale inteligenţei, este posibil şi ca o specie să reprezinte o ameninţare existenţială pentru alta, chiar dacă celei din urmă îi lipseşte aprecierea pentru muzică şi literatură.*

Soluţii

Putem oare să abordăm avertizările lui Wiener direct? Putem proiecta sisteme de inteligenţă artificială ale căror scopuri nu intră în conflict cu ale noastre, pentru a putea fi mulţumiţi de modul în care se comportă? La prima vedere, pare fără speranţă, pentru că se va dovedi cu siguranţă nerealist să ne scriem corect scopurile sau să ne imaginăm toate modurile contraintuitive prin care o entitate superinteligentă le poate atinge.

Dacă tratăm sistemele de inteligenţă artificială superinteligentă ca şi cum ar fi cutii negre venite din spaţiu, atunci, într-adevăr, nu mai avem nicio speranţă. În schimb, abordarea pe care se pare că suntem obligaţi să o folosim, dacă e să avem orice fel de încredere în rezultat,

este să definim o problemă formală *F* şi să proiectăm sisteme de inteligenţă artificială care să poată rezolva probleme de tipul *F*, astfel încât, indiferent de cât de bine rezolvă sistemul problemele *F*> noi vom fi mulţumiţi în mod garantat cu soluţia. Dacă putem să concepem un *F* corespunzător care să aibă această proprietate, vom putea să creăm o inteligenţă artificială *despre care putem dovedi că este benefică*.

Iată un exemplu pentru cum *nu* ar trebui să procedăm: fie ca o recompensă de valoare scalară să fie furnizată periodic de către om maşinii, în funcţie de cât de bine s-a comportat maşina în perioada respectivă, şi fie *F* problema maximizării sumei de recompense aşteptate de maşină. Soluţia optimă la această problemă nu este, aşa cum ar putea spera unii, să te comporţi bine, ci să preiei controlul asupra omului şi să-l forţezi să-ţi ofere un flux de recompense maxime. Aceasta este cunoscută drept problema *wireheadingy* bazată pe observaţii conform cărora oamenii înşişi sunt susceptibili să întâmpine aceeaşi problemă dacă centrele plăcerii din creier le sunt stimulate electronic.

Există, consider eu, o abordare care ar putea să funcţioneze. Oamenii pot fi descrişi în mod rezonabil ca având preferinţe (de cele mai multe ori implicite) pentru vieţile lor viitoare – adică, dacă i se acordă suficient timp şi ajutoare vizuale nelimitate, omul poate exprima o preferinţă (sau indiferenţă) când i se oferă alegerea între două vieţi viitoare care i se deschid în toate aspectele. (Această idealizare ignoră posibilitatea ca minţile noastre să fie compuse din subsisteme cu preferinţe incompatibile; dacă este adevărat, acest lucru ar limita abilitatea unei maşini de a ne satisface optim preferinţele, dar nu pare să ne împiedice să proiectăm maşini ce evită rezultate catastrofale.) Problema formală *F* care trebuie rezolvată de maşină în acest caz este maximizarea preferinţelor vieţii viitoare a omului, neştiind iniţial cu siguranţă care sunt acestea. În plus, deşi preferinţele vieţii viitoare sunt variabile ascunse, sunt înrădăcinate într-o sursă voluminoasă de dovezi – şi anume, toate alegerile omeneşti făcute vreodată. Această formulă evită problema lui Wiener: maşina poate învăţa mai multe despre preferinţele umane în timp, bineînţeles, dar nu va atinge niciodată certitudinea.

O definiţie mai precisă este dată de cadrul învăţării cooperative cu întărire inversă[14] sau CIRL. O problemă CIRL implică doi agenţi: unul uman şi celălalt robot. Deoarece există doi agenţi, problema este ceea ce economiştii numesc un joc. Este un joc de informare parţială, deoarece omul cunoaşte funcţia de recompensare, iar robotul nu – chiar dacă funcţia robotului este să o maximizeze.

Un exemplu simplu: să presupunem că lui Harriet, omul, îi place să colecteze agrafe de hârtie şi capse şi că funcţia ei de recompensare depinde de cât de multe are din fiecare. Mai exact, dacă are *p* agrafe şi

s capse, gradul său de fericire este unde este, în esenţă, un curs de schimb între agrafele de hârtie şi capse. Dacă este 1, înseamnă că-i plac numai agrafele de hârtie; dacă este 0, îi plac doar capsele; dacă este 0,5, îi plac ambele în aceeaşi măsură şi aşa mai departe. Este sarcina lui Robby, robotul, să producă agrafele de hârtie şi capsele. Obiectivul jocului este că Robby vrea să o facă pe Harriet fericită, dar nu cunoaşte valoarea lui aşa că nu ştie sigur cât de multe din fiecare trebuie să producă.

Iată cum funcţionează jocul. Fie valoarea reală a lui 0,49 – adică Harriet are o mică preferinţă pentru capse în detrimentul agrafelor de hârtie. Şi să presupunem că Robby are o bănuială anterioară uniformă despre – adică, crede că poate fi în egală măsură orice valoare între 0 şi 1. Harriet face acum o mică demonstraţie, producând două agrafe, două capse sau câte una din fiecare. Apoi, robotul poate produce nouăzeci de agrafe, nouăzeci de capse sau cincizeci din fiecare. Ai putea crede că Harriet, care preferă capsele, ar trebui să producă două capse. Dar, în acest caz, răspunsul natural al lui Robby ar fi să producă nouăzeci de capse (cu o valoare totală pentru Harriet de 49, 5), care este un rezultat mai puţin dorit pentru Harriet decât cincizeci din fiecare (valoare totală 50,0). Soluţia optimă a acestui joc este ca Harriet să producă câte una din fiecare, pentru ca astfel Robby să producă câte cincizeci din fiecare. Astfel, modul în care este definit jocul o încurajează pe Harriet să-l „înveţe” pe Robby – atât timp cât ştie că Robby o urmăreşte cu atenţie.

În cadrul CIRL, se poate formula şi rezolva problema comutatorului de oprire – adică, problema metodei prin care poţi împiedica robotul să-şi dezactiveze comutatorul de oprire. (Turing poate răsufla liniştit.) Un robot care este nesigur de preferinţele omului beneficiază de pe urma faptului că este oprit, pentru că înţelege că omul va apăsa pe comutatorul de oprire pentru a împiedica robotul să facă ceva contrar acelor preferinţe. Astfel robotul este stimulat să păstreze comutatorul de oprire şi stimulentul său derivă direct din nesiguranţa sa cu privire la preferinţele umane.[15]

Exemplul comutatorului de oprire sugerează şabloane pentru designuri cu agent controlabil şi pune la dispoziţie cel puţin un caz de sistem despre care se poate dovedi că este benefic în sensul prezentat mai sus. Abordarea generală seamănă cu problemele de proiectare a mecanismelor din economie, în care o persoană stimulează alţi agenţi să se comporte în moduri benefice pentru proiectant. Diferenţa-cheie de aici este că noi construim unul dintre agenţi pentru a servi celuilalt.

Există motive să credem că această abordare poate funcţiona practic. În primul rând există informaţii abundente scrise şi filmate despre oameni făcând lucruri (şi despre alţi oameni care

reacţionează). Tehnologia de construire a modelelor de preferinţe umane din acest depozit va fi probabil disponibilă cu mult timp înainte de crearea sistemelor de inteligenţă artificială superinteligentă. În al doilea rând, există stimulente economice pe termen scurt pentru ca roboţii să înţeleagă preferinţele umane: dacă un robot domestic proiectat prost găteşte pisica la cină, fără să-şi dea seama că valoarea ei sentimentală este mai mare decât valoarea ei nutriţională, industria roboţilor domestici va da faliment.

Însă există dificultăţi clare cu o abordare care se aşteaptă de la robot să înveţe preferinţele inerente din comportamentul omului. Oamenii sunt iraţionali, inconsecvenţi, au voinţa slabă şi sunt limitaţi din punctul de vedere al puterii de calcul, aşa că acţiunile lor nu reflectă întotdeauna adevăratele lor preferinţe. (Gândeşte-te, de exemplu, la doi oameni care joacă şah. De obicei, unul dintre ei pierde, dar nu dinadins!) Aşadar roboţii pot învăţa din comportamentul uman neraţional doar cu ajutorul unor modele cognitive ale oamenilor mult mai bune. În plus, restricţiile practice şi sociale vor împiedica satisfacerea la maximum a tuturor preferinţelor simultan, ceea ce înseamnă că roboţii trebuie să medieze între preferinţele conflictuale – lucru cu care filosofii şi cercetătorii sociali se luptă de milenii întregi. Şi ce ar trebui să înveţe roboţii de la oamenii care se bucură de suferinţa altora? Probabil este mai bine să eliminăm astfel de preferinţe din calculele roboţilor.

Găsirea unei soluţii la problema controlării inteligenţei artificiale este o sarcină importantă; poate fi, în cuvintele lui Bostrom, „sarcina esenţială a vremurilor noastre”. Până acum cercetarea inteligenţei artificiale s-a concentrat pe sisteme care sunt mai bune la luarea deciziilor, dar acest lucru este diferit de a lua decizii mai bune. Indiferent cât de bună este maximizarea unui algoritm şi indiferent cât de precis este modelul său de lume, deciziile unei maşini pot fi inefabil de stupide în ochii unui om obişnuit dacă funcţia sa utilitară nu este aliniată cu valorile umane.

Această problemă necesită o modificare a definiţiei inteligenţei artificiale înseşi – de la un domeniu care se preocupă de inteligenţa pură, independentă de obiectiv, la un câmp care se preocupă cu sisteme care sunt vădit benefice pentru oameni. Dacă luăm în serios această problemă, probabil vom găsi modalităţi noi de a ne gândi la inteligenţa artificială, la scopul ei şi la relaţia noastră cu ea.

Capitolul 4
A treia lege

George Dyson

George Dyson este un istoric al ştiinţei şi tehnologiei şi autorul cărţilor Baidarka: The Kayak, Darwin Among the Machine Sy Project Orion şi Turings Cathedral.

În 2005, George Dyson, istoric al ştiinţei şi tehnologiei, a vizitat Google la invitaţia unor ingineri de-acolo. Ocazia a fost reprezentată de cea de-a şaizecea aniversare a propunerii lui John von Neumann de computer digital. În urma acestei vizite, George a scris un eseu, „Catedrala lui Turing", care avertiza publicul pentru prima dată despre ce-şi propun să facă fondatorii Google. „Nu scanăm toate aceste cărţi pentru a putea fi citite de oameni", îi explica una dintre gazde după prelegere. „Le scanăm pentru a fi citite de o inteligenţă artificială".

George oferă o poveste alternativă pentru era digitală. Interesele sale au inclus dezvoltarea lui caiacului aleutin, evoluţia tehnicii de calcul digitale şi a telecomunicaţiilor, originile universului digital şi o cale nestrăbătută către spaţiu. Cariera sa (nu a terminat niciodată liceul, însă a primit un titlu onorific de doctor de la Universitatea

Victoria) s-a dovedit imposibil de clasificat şi cărţile sale la fel.

Îi place să evidenţieze că tehnica de calcul analogică, despre care se credea cândva că este depăşită, la fel ca analizorul diferenţial, a revenit. Argumentează că, deşi folosim componente digitale, la un anumit moment, tehnica de calcul analogică efectuată de sistem depăşeşte cu mult complexitatea codului digital cu care este construit. Consideră că inteligenţa artificială adevărată – cu sistemele analogice de control care iau naştere dintr-un substrat digital, la fel cum computerele digitale au luat naştere din componentele analogice după cel de-al Doilea Război Mondial – nu este atât de îndepărtată pe cât credem.

În acest eseu, George contemplă diferenţa dintre calculele analogice şi cele digitale şi constată că analogicul este bine şi sănătos în lumea contemporană. Răspunsul naturii la încercarea de a programa maşinile să controleze totul poate fi dat sub forma unor maşini fără programare asupra cărora nimeni nu are control.

Istoria tehnicii de calcul poate fi împărţită într-un Vechi Testament şi un Nou Testament: înainte şi după ce computerele digitale electronice şi codurile cărora le-au dat naştere au proliferat pe Pământ. Profeţii Vechiului Testament, care ne-au pus la dispoziţie logica elementară, i-au inclus pe Thomas Hobbes şi Gottfried Wilhelm Leibniz. Profeţii Noului Testament iau înclus pe Alan Turing, John von Neumann, Claude Shannon şi Norbert Wiener. Aceştia ne-au pus la dispoziţie maşinile.

Alan Turing se întreba ce ar face o maşină inteligentă. John von

Neumann se întreba ce ar putea face ca maşinile să se autoreproducă. Claude Shannon se întreba ce ar putea face ca maşinile să comunice fiabil, indiferent de zgomotul care ar interveni. Norbert Wiener se întreba cât timp mai avem până când maşinile vor prelua controlul.

Avertizările lui Wiener cu privire la sistemele de control care se prelungesc dincolo de controlul oamenilor au apărut în 1949, când a fost introdusă prima generaţie de computere digitale electronice cu programe stocate. Sistemele necesitau o supraveghere directă din partea programatorilor umani, subminând motivele sale de îngrijorare. Care este problema atât timp cât programatorii controlează maşinile? De atunci dezbaterea cu privire la riscurile controlului autonom a rămas asociată cu dezbaterea cu privire la puterile şi limitările maşinilor codate digital. În ciuda puterilor lor uluitoare, a fost observată foarte puţină autonomie reală. Aceasta este o presupunere periculoasă. Dacă puterea tehnicii de calcul digitale este înlocuită de altceva?

Electronica a trecut prin două tranziţii fundamentale în ultima sută de ani: de la analogic la digital şi de la tuburi vidate la starea solidă. Faptul că aceste tranziţii au apărut împreună nu înseamnă că sunt legate inextricabil. La fel cum calculul digital a fost implementat cu ajutorul componentelor cu tuburi vidate, calculul analogic poate fi implementat în starea solidă. Calculul analogic este bine sănătos, chiar dacă tuburile vidate au dispărut din comerţ.

Nu există nicio distincţie precisă între calculul analogic şi cel digital. În general, calculul digital lucrează cu numere întregi, secvenţe binare, logică deterministă şi timp idealizat în increment discrete, în timp ce calculul analogic lucrează cu numere reale, logică nedeterministă şi funcţii continue, inclusiv timpul aşa cum există el ca un continuum în lumea reală.

Imaginează-ţi că trebuie să găseşti mijlocul unui drum. Poţi să-i măsori lăţimea cu orice increment disponibil şi apoi poţi calcula digital mijlocul celui mai apropiat increment. Sau poţi folosi o sfoară drept computer analogic, mapând lăţimea drumului cu lungimea sforii şi găsind mijlocul acesteia, fără a fi limitat la increment, unind capetele sforii şi apoi întinzând-o.

Multe sisteme funcţionează atât în regim analogic, cât şi în regim digital. Un copac integrează o gamă variată de intrări ca funcţii continue, dar, dacă tai acel copac, observi că a numărat digital anii tot timpul.

În calculul analogic, complexitatea rezidă în topologia reţelei, şi nu în cod. Informaţia este prelucrată ca funcţii continue de valori, precum tensiunea şi frecvenţa relativă a pulsului, şi nu prin operaţii logice pe şiruri discrete de biţi. Calculul digital, intolerant la erori sau ambiguităţi, depinde de corectarea erorilor în orice pas pe care-l face.

Calculul analogic tolerează erori, permiţându-ţi să trăieşti cu ele.

Natura foloseşte codarea digitală pentru depozitarea, replicarea şi recombinarea secvenţelor de nucleotide, dar se bazează pe calcul analogic, funcţionând pe sisteme nervoase, pentru inteligenţă şi control. Sistemul genetic din toate celulele vii este un computer cu program stocat. Creierele nu sunt.

Computerele digitale execută transformări între două specii de biţi: biţi ce reprezintă diferenţele în spaţiu şi biţi ce reprezintă diferenţele în timp. Transformările dintre aceste două forme de informaţie, secvenţă şi structură, sunt guvernate de programarea computerului şi, atât timp cât computerele au nevoie de programatori umani, noi păstrăm controlul.

Computerele analogice mediază şi transformările dintre două forme de informaţii: structura în spaţiu şi comportamentul în timp. Nu există nici cod, nici programare. Cumva – şi nu înţelegem prea bine cum – natura a dezvoltat computere analogice cunoscute sub numele de sisteme nervoase, care înglobează informaţiile absorbite din lume. Învaţă. Unul dintre lucrurile pe care le învaţă este controlul. Învaţă să-şi controleze propriul comportament şi învaţă să-şi controleze mediul în măsura în care pot.

Informatica are o istorie îndelungată – care se întinde chiar înainte de *existenţa* informaticii – de implementare a reţelelor neuronale, dar majoritatea au fost simulări de reţele neuronale făcute de computere digitale, şi nu reţele neuronale aşa cum le-a creat natura. Acest lucru începe să se schimbe, de jos în sus, deoarece factorul întreit al războiului cu drone, vehiculelor autonome şi telefoanelor celulare împinge dezvoltarea microprocesoarelor neuromorfe care implementează reţele neuronale efective, şi nu simulări ale reţelelor neuronale direct în siliciu (şi alte substraturi potenţiale); şi de sus în jos, deoarece cele mai mari şi mai de succes întreprinderi ale noastre se întorc din ce în ce mai mult la calculele analogice, încercând să se infiltreze şi să controleze lumea.

În timp ce dezbatem inteligenţa computerelor digitale, calculul digital se furişează tăcut peste cel digital, în acelaşi fel în care componentele analogice precum tuburile vidate au fost restructurate pentru a construi computerele digitale în urma celui de-al Doilea Război Mondial. Procesoarele deterministe individuale în stare finită, ce rulează coduri finite, formează organisme metazoare în stare nonfinită, nedeterministe şi la scară largă care zburdă libere în lumea reală. Sistemele hibride analogice şi digitale rezultate tratează colectiv fluxurile de biţi, la fel cum este fluxul de electroni într-un tub vidat, nu individual, cum sunt trataţi biţii de dispozitivele în stare discretă care generează fluxul. Biţii sunt noii electroni. Analogicul a revenit şi natura sa este să preia controlul.

Guvernând totul de la fluxul bunurilor până la fluxul traficului şi al ideilor, aceste sisteme funcţionează statistic, deoarece informaţiile codate cu frecvenţă de puls sunt prelucrate într-un neuron sau într-un creier. Emergenţa inteligenţei atrage atenţia lui *Homo sapiens* dar ar trebui să ne îngrijorăm mai mult de emergenţa controlului.

Imaginează-ţi că eşti în anul 1958 şi încerci să aperi partea continentală a Statelor Unite ale Americii de atacurile aeriene. Pentru a distinge aeronavele ostile, unul dintre lucrurile de care ai nevoie, în afară de o reţea de computere şi radare cu avertizare în timp util, este o hartă cu traficul aerian comercial, actualizată în timp real. Statele Unite au construit o astfel de hartă şi au numit-o SAGE (Semi Automatic Ground Environment). SAGE a dat la rândul său naştere proiectului Sabre, primul sistem integrat de rezervare a zborurilor în timp real. Sabre şi urmaşii săi au devenit în curând nu doar o hartă cu locurile disponibile, ci şi un sistem care a început să controleze, cu o inteligenţă descentralizată, unde vor zbura companiile aeriene şi când.

Dar nu este oare pe undeva o cameră de control, cu cineva la comenzi? Poate că nu. Să zicem, de exemplu, că vom construi un sistem care să cartografieze traficul de pe autostrăzi în timp real, pur şi simplu dând maşinilor acces la acea hartă în schimbul informaţiilor despre viteza şi localizarea lor la momentul respectiv. Rezultatul este un sistem de control complet descentralizat. Nicăieri nu se află un model de control al sistemului în afara sistemului însuşi.

Imaginează-ţi că este primul deceniu din secolul al XXI-lea şi vrei să urmăreşti complexitatea relaţiilor umane în timp real. Pentru viaţa socială din cadrul unei facultăţi mici, ai putea construi o bază de date centrală, încercând să o ţii actualizată, dar întreţinerea ei ar deveni copleşitoare dacă sistemul ar fi implementat la o scară mai largă. Mai bine să distribui copii gratuite ale unui cod semiautonom simplu, găzduit local, şi să laşi reţeaua socială să se actualizeze singură. Codul este executat de computere digitale, dar calculele analogice efectuate de sistem ca întreg depăşesc cu mult complexitatea codului pe care îl au la bază. Modelul rezultat, codat cu frecvenţa pulsului, al graficului social *devine* graficul social. Se întinde prin campus şi apoi în lume.

Dar dacă ai vrea să construieşti o maşină care să surprindă ce *înseamnă* tot ce e cunoscut de specia umană? Cu Legea lui Moore în spate, nu durează prea mult să digitalizezi toate informaţiile din lume. Scanezi toate cărţile tipărite vreodată, colectezi toate e-mailurile scrise vreodată şi strângi patruzeci şi nouă de ani de imagini video la fiecare douăzeci şi patru de ore, în timp ce urmăreşti unde se află oamenii şi ce fac, în timp real. Dar cum surprinzi *sensur*

Chiar şi în epoca tuturor lucrurilor digitale, acesta nu poate fi

definit cu o stricteţe logică, deoarece sensul, la oameni, nu este fundamental logic. După ce ai colectat toate răspunsurile posibile, nu poţi să faci decât să inviţi întrebări bine definite şi să compilezi o hartă ponderată cu frecvenţe de puls pentru felul în care se conectează toate lucrurile între ele. Foarte repede, sistemul tău nu va observa şi mapa doar sensul lucrurilor, va începe şi să *construiască* sens. În timp, va *controla sens*ul, în acelaşi fel în care hărţile traficului încep să controleze fluxul traficului chiar dacă nimeni nu pare să fie la control.

Există trei legi ale inteligenţei artificiale.

Prima, cunoscută sub numele de Legea lui Ashby, după cibernetician W. Ross Ashby, autorul cărţii *Design for a Brain*, spune că sistemul de control efectiv trebuie să fie la fel de complex ca sistemul pe care-l controlează.

A doua lege, articulată de John von Neumann, spune că trăsătura definitorie a unui sistem complex este că îşi constituie singur cea mai simplă descriere comportamentală. Cel mai simplu model complet al unui organism este organismul însuşi. Dacă încerci să reduci comportamentul sistemului la o descriere formală, îngreunezi lucrurile, nu le uşurezi.

A treia lege spune că orice sistem suficient de simplu să poată fi înţeles nu va fi suficient de complicat pentru a se comporta inteligent şi orice sistem suficient de complicat pentru a se comporta inteligent va fi prea complicat de înţeles.

A treia lege îi linişteşte pe cei care cred că, până când vom înţelege inteligenţa, nu trebuie să ne facem probleme legate de naşterea inteligenţei superomeneşti în maşini. Dar există o portiţă în cea de-a treia lege. Este complet posibil să construieşti ceva fără să-l înţelegi. Nu trebuie să înţelegi complet cum funcţionează un creier pentru a construi unul care funcţionează. Aceasta este o portiţă pe care nicio supraveghere a algoritmilor de către programatori şi consultanţii lor etici, oricât ar fi ea de strictă, nu o poate închide. Inteligenţa artificială ostentativ „bună” este un mit. Relaţia noastră cu inteligenţa artificială adevărată va ţine întotdeauna de încredere, nu de dovadă.

Ne facem prea multe probleme legate de inteligenţa maşinilor şi nu ne îngrijorăm suficient de autoreproducere, comunicare şi control. Următoarea revoluţie în puterea de calcul va fi semnalizată de ascensiunea sistemelor analogice, asupra cărora programarea digitală nu mai are niciun control. Răspunsul naturii pentru cei care consideră că pot construi maşini care să controleze totul va fi să le permită să construiască o maşină care-i va controla pe ei în schimb.

Capitolul 5

Daniel C. Dennett

Daniel C. Dennett este profesor universitar şi profesor Austin B. Fletcher de filosofie şi coadministrator al Centrului pentru Studii Cognitive de la Universitatea Tuft. Este autorul a douăsprezece cărţi, inclusiv Consciousness Explained şi, mai recent, From Bacteria to Bach and Back: the Evolution of Minds.

Dan Dennett este filosoful preferat al comunităţii inteligenţei artificiale. Este cunoscut în ştiinţa cognitivă probabil pentru conceptul său de sisteme intenţionale şi modelul său pentru conştiinţa umană, care schiţează o arhitectură computaţională pentru realizarea fluxului conştiinţei în cortexul cerebral puternic paralelizat. Acest computaţionalism fără compromisuri a primit critici puternice din partea unor filosofi precum John Searle, David Chalmers şi răposatul Jerry Fodor, care au spus că cele mai importante aspecte ale conştiinţei – intenţionalitatea şi qualia subiectivă – nu pot fi calculate.

Cu douăzeci şi cinci de ani în urmă, îl vizitam pe Marvin Minsky, unul dintre pionierii inteligenţei artificiale, şi l-am întrebat de Dan. „Este cel mai bun filosof al nostru din prezent – următorul Bertrand Russeli, a spus Marvin, adăugând că, spre deosebire de filosofii tradiţionali, Dan studia neuroştiinţa, lingvistica, inteligenţa artificială, informatica şi psihologia. „Redefineşte şi reformează rolul filosofului. Bineînţeles, Dan nu înţelege teoria mea despre societatea minţii, dar nimeni nu este perfect".

Perspectiva lui Dan asupra eforturilor depuse de cercetătorii din domeniul inteligenţei artificiale pentru a crea inteligenţe artificiale superinteligente arată întotdeauna o judecată sănătoasă. Ce, să-mi fac probleme eu? În acest eseu, ne reaminteşte că inteligenţele artificiale, mai presus de toate, trebuie privite – şi tratate – ca nişte instrumente, şi nu ca pe nişte colegi umani.

A fost interesat de teoria informaţiei încă din zilele sale de studenţie de la Oxford. De fapt, mi-a spus că, încă de la începutul carierei sale, era foarte interesat să scrie o carte despre ideile cibernetice ale lui Wiener. În calitate de gânditor care îmbrăţişează metoda ştiinţifică, din charisma sa face parte şi disponibilitatea de a greşi. Despre un articol recent intitulat „Ce este informaţia?", a anunţat: „Îl susţin, dar este în revizuire. Deja am trecut dincolo de el şi mi-am dat seama că există o cale mai bună de a aborda câteva dintre aceste probleme". Probabil va rămâne calm şi raţional cu privire la subiectul cercetării inteligenţei artificiale, deşi a recunoscut, adesea, că propriile sale idei evoluează – la fel cum

ar trebui să facă ideile tuturor.

Mulţi au reflectat asupra ironiei de a citi o carte mare când eşti prea tânăr să o apreciezi. Să pui un clasic în stiva de cărţi *deja citite*, izolându-te astfel de alte influenţe din partea lui, spicuind doar câteva idei prost înţelese din el, este o reţetă pentru neglijenţă care este rareori benignă. Această idee m-a lovit puternic când am citit *Folosirea omului de către om* la peste şaizeci de ani după întâlnirea mea juvenilă cu această carte. Ar trebui să ne facem cu toţii un obicei regulat să recitim cărţi din tinereţe, unde putem descoperi previzualizări clare ale câtorva dintre „descoperirile" şi „invenţiile" noastre ulterioare, alături de o bogăţie de observaţii la care eram destinaţi să fim insensibili până când minţile noastre nu se vor coace, antrena şi spori prin confruntarea cu problemele vieţii.

Scriind într-o vreme în care tuburile vidate încă erau elementele de fundaţie ale electronicii şi existau doar câteva computere, Norbert Wiener şi-a imaginat viitorul pe care-l trăim acum în detalii impresionante şi cu destul de puţine greşeli evidente. Faimosul articol din 1950 al lui Alan Turing „Maşinile de calcul şi inteligenţa", apărut în revista de filosofie *Mind*, a prevăzut dezvoltarea inteligenţei artificiale, la fel a făcut şi Wiener, însă Wiener a văzut mai departe şi mai adânc, recunoscând că inteligenţa artificială nu va imita – şi înlocui – pur şi simplu fiinţele umane în multe activităţi inteligente, ci va şi schimba oamenii pe parcurs:

> *„Nu suntem decât vârtejuri într-un râu permanent. Nu suntem ceva care să continue, ci tipare care se perpetuează"*[16]

Când au fost scrise aceste cuvinte, puteau fi respinse cu uşurinţă ca altă exagerare heracliteană. Da, da, nu poţi să păşeşti în acelaşi râu de două ori. Dar conţine germenii revoluţiei în perspectivă. Astăzi ştim cum să ne raportăm la sisteme adaptive complexe, atractori bizari, minţi extinse şi homeostază, o schimbare de perspectivă care promite să şteargă cu buretele „discontinuitatea explicativă"[17] dintre minte şi mecanism, spirit şi materie, discontinuitate care este încă apărată cu ferocitate de cartezienii moderni care nu pot suporta ideea că noi – noi *înşine* – suntem tipare de materie purtătoare de informaţie care se autoperpetuează, şi nu „ceva care să continue". Acele tipare sunt remarcabil de rezistente şi autorestauratorii, dar în acelaşi timp exploatatori versatili, oportunişti şi egoişti ai oricărei ocazii noi şi disponibile pentru valorificare în goana lor după perpetuare. Şi aici devine riscant, spunea Wiener. Când oportunităţile atrăgătoare abundă, suntem deschişi să plătim puţin şi să acceptăm costuri de afaceri mici, chiar triviale, pentru a accesa puteri noi. Şi devenim rapid atât de dependenţi de instrumentele noastre, încât pierdem capacitatea de a prospera fără ele. Opţiunile devin obligatorii.

Este o poveste foarte veche, cu multe capitole bine-cunoscute în istoria evoluţionistă. Majoritatea mamiferelor îşi pot sintetiza propria vitamina C, dar primatele, alegând o dietă ce conţine în mare parte fructe, au pierdut această abilitate înnăscută. Suntem acum obligaţi să ingurgităm vitamina C, dar nu suntem obligaţi să fim frugivori ca verii noştri, primatele, pentru că am optat pentru o tehnologie care ne permite să facem şi să luăm vitaminele atunci când avem nevoie. Tiparele autoperpetuante pe care le numim oameni sunt dependente astăzi de haine, alimente gătite, vitamine, vaccinuri, carduri de credit, telefoane inteligente şi internet. Şi – mâine, dacă nu deja de astăzi – de inteligenţa artificială.

Wiener a prevăzut problemele pe care Turing şi ceilalţi optimişti le-au trecut în mare cu vederea. Pericolul real, a spus el, este

> *„Ca aceste maşini, deşi neajutorate pe cont propriu, să poată fi utilizate de om sau o organizaţie de oameni pentru a-şi spori controlul asupra restului rasei sau că liderii politici ar putea încerca să controleze populaţiile nu neapărat prin intermediul maşinilor, ci prin tehnici politice înguste şi indiferente la posibilităţile umane, ca şi cum ar fi fost, de fapt, concepuţi mecanic”.*[18]

Puterea, recunoştea el, rezidă în primul rând în algoritmi, nu în componentele hardware pe care rulează, deşi hardware-ul de astăzi face posibili algoritmi care ar fi părut ridicol de greoi pe vremea lui Wiener. Ce putem spune despre aceste „tehnici” care sunt „înguste şi indiferente la posibilităţile omeneşti”? Au fost prezentate de multe ori, unele evident benefice, altele evident periculoase, şi multe în omniprezentul teren conflictual al controversei.

Gândiţi-vă la câteva dintre dispute. Prietenului meu decedat Joe Weizenbaum, succesorul lui Wiener în calitatea de profet Ieremia de la MIT în domeniul hi-tech, îi plăcea să observe despre cardurile de credit că, oricare ar fi punctele lor forte, ofereau de asemenea şi o modalitate ieftină şi sigură pentru guvern şi corporaţii să urmărească deplasările, obiceiurile şi dorinţele oamenilor. Anonimitatea numerarului a fost mult subapreciată, în afară de traficanţii de droguri şi alţi infractori, şi acum este pe cale de dispariţie. Acest lucru poate îngreuna spălarea de bani pe viitor, însă programele de căutare ale tiparelor folosite de inteligenţa artificială contra acestei probleme au efectul secundar de a ne face pe toţi mai transparenţi la un „bloc de fiinţe umane” care ar putea „încerca să ne controleze”.

În ceea ce priveşte artele, inovaţia înregistrărilor digitale audio şi video ne permite să plătim un preţ mic (în ochii tuturor, dar cel mai arzător, în ochii audiofililor şi cinefililor) când abandonăm formatele analogice şi, în schimb, oferă o reproducere uşoară – mult prea uşoară? – a operelor de artă cu o fidelitate aproape perfectă. Dar

costul ascuns este uriaş. Ministerul Adevărului imaginat de Orwell este acum o posibilitate practică. Tehnici ale inteligenţei artificiale de creare a falsurilor aproape nedetectabile pentru „înregistrările" de întâlniri devin acum disponibile, lucru care va face instrumentele de investigare pe care am ajuns să le luăm de-a gata în ultimii 150 de ani să devină depăşite. Vom abandona scurta epocă a dovezilor fotografice şi ne vom întoarce la lumea de dinainte, a memoriei umane şi a încrederii date de standardul aurului, sau vom dezvolta tehnici noi de apărare şi atac în cursa pentru înarmare a adevărului? (Ne putem imagina o întoarcere la filmul *analog* expus la lumină, păstrat în sisteme „cu protecţie la acces" până când vor fi arătate juriilor etc., dar cât de mult va dura până când cineva găseşte o soluţie să infecteze aceste sisteme cu îndoială? Una dintre lecţiile grele ale experienţelor recente este că sarcina distrugerii unei reputaţii pentru credibilitate este mult mai ieftină decât sarcina protejării aceleiaşi reputaţii.) Wiener a văzut acest fenomen în cel mai general aspect al său: „pe termen lung, nu mai există nicio diferenţă între a ne înarma noi înşine sau a ne înarma duşmanii". Era informaţiei este în acelaşi timp şi era dezinformării.

Ce putem face? Trebuie să ne regândim priorităţile cu ajutorul analizelor eronate, dar pline de pasiune ale lui Wiener, Weizenbaum şi ceilalţi critici serioşi ai tehnofiliei noastre. O frază-cheie mi se pare observaţia aproape improvizată de mai sus că „aceste maşini sunt aproape neajutorate pe cont propriu". După cum am argumentat recent, construim instrumente, nu colegi, şi marele pericol este să nu apreciem diferenţa, pe care ar trebui să încercăm să o evidenţiem, marcând-o şi apărând-o cu inovaţii politice şi legale.

Poate că cea mai bună cale de a vedea ce ratăm este să ne aducem aminte că Alan Turing însuşi a suferit de o lipsă de imaginaţie complet de înţeles când a formulat faimosul său test Turing. După cum ştie toată lumea, este o adaptare a „jocului său de imitaţie" în care un om, care nu se vede, comunică verbal cu un judecător, încercă să-l convingă că este, de fapt, o femeie, în timp ce o femeie, de asemenea ascunsă şi care comunică cu judecătorul, încearcă să-l convingă pe judecător că ea este femeia. Turing credea că aceasta va fi o provocare grea pentru un bărbat (sau pentru o femeie care se preface că este bărbat), exploatând o bogăţie de cunoştinţe despre cum gândeşte şi reacţionează celălalt sex, despre ce are tendinţa să favorizeze şi să ignore. Cu siguranţă (*hopa*!) orice bărbat care poate învinge o femeie la a fi perceput ca fiind femeie ar fi un agent inteligent. Ce nu a prevăzut Turing este puterea inteligenţei artificiale cu învăţare profundă de a obţine această bogăţie de informaţii într-o formă exploatabilă *fără să fie nevoie să o înţeleagă*. Turing s-a gândit la un agent perspicace şi plin de imaginaţie (aşadar, conştient) care-şi

gândea isteţ răspunsurile în baza „teoriei” detaliate despre ce este probabil să facă şi să spună o femeie. Design inteligent descendent, pe scurt. Cu siguranţă nu s-a gândit că un bărbat care câştigă jocul imitaţiei va *deveni* cumva femeie; îşi imagina că va exista tot timpul conştiinţa bărbatului la baza acestui spectacol. Premisa ascunsă din argumentul lui Turing era: doar un *agent* conştient inteligent poate dezvolta şi controla o strategie victorioasă în jocul imitaţiei. Aşadar, era convingător pentru Turing (şi pentru alţii, printre care mă număr şi eu, un dârz apărător al testului lui Turing) să argumenteze că „o maşină de calcul” care ar putea trece drept om într-o întrecere cu un om nu este neapărat conştientă aşa cum este omul, dar trebuie să fie un soi de *agent* conştient. Cred că este o poziţie care se poate apăra – singura poziţie care se poate apăra –, dar trebuie să înţelegi cât de perspicace şi de ingenios ar trebui să fie un judecător pentru a expune superficialitatea faţadei pe care ar putea-o prezenta o inteligenţă artificială cu învăţare profundă (un instrument, nu un coleg).

Ce nu a prevăzut Turing este capacitatea nemaivăzută a computerelor foarte rapide de a cerne fără discernământ prin seturile mari de date, pe care internetul le pune la dispoziţie în cantităţi nesfârşite, găsind tipare probabilistice în activităţile oamenilor care pot fi utilizate pentru a improviza răspunsuri care să pară „autentice” în dispozitivul de ieşire pentru aproape orice probă pe care un judecător s-ar gândi s-o administreze. Şi Wiener a subestimat această posibilitate, ţinând cont de faptul că slăbiciunea care trădează maşina este să nu fie capabilă să

> *„la în considerare vastul interval de probabilitate ce caracterizează situaţia umană”*[19]

Dar a lua în considerare acest vast interval de probabilitate este exact punctul forte al inteligenţei artificiale. Singura slăbiciune din armura inteligenţei artificiale este acest cuvânt „vast”; posibilităţile omeneşti, graţie limbajului şi culturii pe care le produce, sunt cu adevărat vaste[20]. Indiferent de cât de multe tipare am putea găsi cu inteligenţa artificială în fluviul de date care se găsesc pe internet, există mult mai multe posibilităţi decât sunt înregistrate aici. Doar o fracţie (dar nu o fracţie nesemnificativă) din înţelepciunea, proiectele, repertoriile şi prostioarele acumulate ale lumii au ajuns pe internet, dar, probabil, o tactică mai bună pentru judecător când confruntă un candidat pentru testul Turing este nu să *caute* aceste elemente, ci să le *creeze* de la zero.

Inteligenţa artificială în manifestările sale curente este parazitară pe inteligenţa umană. Se îndoapă nediferenţiat cu tot ce a fost produs de creatorii umani şi extrage tiparele care pot fi extrase de acolo – inclusiv obiceiurile noastre dăunătoare[21]. Aceste maşini nu au (încă)

obiectivele, strategiile sau capacităţile de autocritică şi inovaţie care să le permită să transcendă bazele lor de date gândindu-se reflexiv la propria lor gândire şi la propriile lor obiective. Sunt, după cum spune Wiener, neajutorate, nu în sensul de agenţi *încătuşaţi* sau agenţi *cu capacităţi reduse*, ci în sensul că nu sunt agenţi deloc – nu au capacitatea de a fi „puse în mişcare de motivele" (cum spunea Kant) care li se prezintă. Este important să păstrăm lucrurile astfel, ceea ceva necesita muncă.

Una dintre erorile din cartea lui Weizenbaum, *Computer Power and Human Reason*, de care am încercat să-l conving în multe ore de discuţii, era că nu se putea hotărî niciodată ce să apere: *inteligenţa artificială este imposibilă*! Sau *inteligenţa artificială este posibilă, dar malefică*! Voia să argumenteze, împreună cu John Searle şi Roger Penrose, că „inteligenţa artificială puternică" este imposibilă, dar nu există argumente solide care să justifice această concluzie. La urma urmei, tot ce ştim acum sugerează că, după cum am mai spus, suntem roboţi făcuţi din roboţi făcuţi din roboţi… până la nivelul proteinelor motorii şi altor elemente asemănătoare, fără ingrediente magice aruncate în amestec pe parcurs. Mesajul mai important şi mai defensibil al lui Weizenbaum era că nu ar trebui să ne străduim să creăm inteligenţă artificială puternică şi ar trebui să fim foarte atenţi în ceea ce priveşte sistemele de inteligenţă artificială pe care le putem crea şi pe care le-am creat deja. După cum te-ai putea aştepta, teza defensibilă este un hibrid: *inteligenţa artificială* (inteligenţa artificială puternică) *este posibilă în principiu, dar nu este de dorit. Inteligenţa artificială care este posibilă din punct de vedere practic nu este neapărat rea decât dacă se confundă cu inteligenţa artificială puternică*!

Intervalul dintre sistemele de astăzi şi sistemele science-fiction ce domină imaginaţia populară este încă uriaş, deşi mulţi oameni – atât nespecialişti, cât şi specialişti – reuşesc să-l subestimeze. Să ne gândim la Watson, de la IBM, care poate fi luat drept un reper de încredere pentru imaginaţiile noastre deocamdată. Este rezultatul unui proiect de cercetare şi dezvoltare la scară foarte largă, întinzându-se peste multe secole pe persoană, după cum notează George Church în aceste pagini, foloseşte de mii de ori mai multă energie decât un creier omenesc (o limitare tehnologică ce, după cum observă tot el, poate fi temporară). Victoria sa în *Jeopardy*! a fost un triumf veritabil, făcut posibil de restricţiile formale ale regulilor emisiunii, dar, pentru a putea concura, chiar şi aceste reguli au trebuit să fie revizuite (un compromis: renunţi la puţină versatilitate, la puţină umanitate şi obţii un spectacol care mulţumeşte mulţimea). Watson nu este o companie plăcută, în ciuda reclamelor amăgitoare de la IBM care sugerează o capacitate conversaţională generală, iar transformarea lui Watson într-un *agent* multidimensional plauzibil ar fi echivalentul transformării

unui calculator de buzunar în Watson. Watson ar putea fi o facultate de bază utilă pentru un astfel de agent, dar mai asemănător unui cerebel sau unei amigdale decât unei minţi – în cel mai bun caz, un subsistem specializat care poate juca un rol de suport mare, dar nici măcar aproape să poată formula scopuri, planuri şi structuri cu perspicacitate în baza experienţelor sale conversaţionale.

De ce am vrea să-l transformăm pe Watson într-un agent gânditor şi creativ? Poate că ideea genială a lui Turing de test operativ ne-a dus într-o capcană: sarcina de a crea cel puţin iluzia unei persoane reale în spatele ecranului, care să ofere o punte peste „valea misterioasă". Pericolul aici este că, de când a lansat Turing această provocare – care era, la urma urmei, o provocare menită să-i *păcălească pe* judecători –, creatorii inteligenţei artificiale au încercat să peticească valea cu artificii omeneşti drăgălaşe, efecte de Disneyficare care-i vor încânta şi dezarma pe neiniţiaţi.

EUZA lui Weizenbaum a fost exemplul principal pentru o astfel de iluzionare superficială şi disperarea sa la uşurinţa cu care programul său rizibil de simplu şi superficial îi putea convinge pe oameni că au o conversaţie candidă este cea care l-a făcut să-şi asume această misiune.

Şi motivele sale de îngrijorare erau întemeiate. Dacă am învăţat ceva din competiţiile restricţionate ale testului Turing pentru Premiul Loebner este că până şi oamenii inteligenţi care nu sunt familiari cu posibilităţile şi scurtăturile programării pe calculator sunt păcăliţi cu uşurinţă de trucuri simple. Atitudinile oamenilor din domeniul inteligenţei artificiale faţă de aceste metode de a disimula în „interfaţa de utilizator" au mers de la dispreţ la celebrare, cu o apreciere generală că aceste trucuri nu sunt profunde, dar pot fi puternice. O schimbare de atitudine foarte binevenită este recunoaşterea candidă că ornamentele omeneşti sunt *reclamă falsă* – ceva de condamnat, nu de aplaudat.

Cum am putea realiza acest lucru? Când vom recunoaşte că oamenii încep să ia decizii de viaţă şi de moarte în mare parte la „sfatul" sistemelor de inteligenţă artificială ale căror operaţii interne sunt de nepătruns în practică, putem vedea un motiv suficient pentru cei care-i încurajează pe oameni în orice fel să acorde mai multă încredere acestor sisteme decât ar trebui să fie traşi la răspundere atât moral, cât şi legal. Sistemele de inteligenţă artificială sunt instrumente foarte puternice – atât de puternice încât până şi experţii au motive întemeiate să nu aibă încredere în propria judecată privind „judecăţile" propriilor lor instrumente. Însă, dacă aceşti utilizatori de instrumente urmează să beneficieze, financiar sau altfel, de folosirea lor în *terra incognita*, trebuie să se asigure că o pot face responsabil, cu un control şi o justificare maxime. Licenţierea şi alocarea operatorilor,

la fel cum îi licenţiem pe farmacişti (şi pe operatorii de macarale!) şi alţi specialişti ale căror greşeli şi erori de judecată pot avea consecinţe grave, îi pot, din cauza presiunii companiilor de asigurări şi a altor asigurători, obliga pe creatorii sistemelor de inteligenţă artificială să facă eforturi extraordinare să caute şi să arate slăbiciuni şi lipsuri în produsele lor şi să-i instruiască pe cei care au dreptul să le folosească.

Ne putem imagina un fel de test Turing invers, în care judecătorul este supus procesului; până când acesta poate detecta slăbiciunea, limitele depăşite, lipsurile dintr-un sistem, nicio licenţă de operare nu va fi emisă. Instruirea mentală necesară pentru a obţine certificarea de judecător va fi dificilă. Îndemnul de a adopta o poziţie intenţională, tactica noastră normală ori de câte ori întâlnim ceea ce pare a fi un agent inteligent, este aproape copleşitor de puternic. Într-adevăr, capacitatea de a rezista tentaţiei de a trata ceea ce pare a fi o persoană ca pe o persoană este un talent urât, care duhneşte a rasism sau specism. Mulţi oameni ar considera cultivarea unei abordări atât de sceptice respingătoare din punct de vedere moral şi putem anticipa că până şi cei mai specializaţi utilizatori de sistem vor cădea pradă ocazional tentaţiei de a se „împrieteni" cu instrumentele lor, fie şi numai pentru a-şi alina disconfortul provocat de realizarea îndatoririlor pe care le au. Indiferent cât de lipsiţi de scrupule ar fi designerii inteligenţei artificiale în deghizarea „umană" a bunurilor produse, ne putem aştepta ca obiceiuri noi de gândire, iniţiative şi artificii conversaţionale, capcane şi blufuri să ia naştere în acest context nou al acţiunii umane. Listele hilar de lungi cu efecte secundare cunoscute ale medicamentelor noi care au reclame la televizor vor fi umbrite de revelaţiile obligatorii ale întrebărilor care *nu pot* primi răspunsuri responsabile din partea unor anumite sisteme, cu pedepse dure pentru cei care „scapă" defecte în produsele lor. Este recunoscut la scară largă că o parte considerabilă din inechitatea economică în creştere de astăzi se datorează bogăţiei acumulate de antreprenorii digitali; trebuie să introducem legi care limitează buzunarele lor adânci pentru binele public. Câteva dintre cele mai adânci buzunare sunt în mod voluntar în opoziţie cu aceste obligaţii de a servi societatea în primul rând şi de a câştiga bani în al doilea rând, dar nu ar trebui să ne bizuim doar pe buna credinţă.

Nu avem nevoie de agenţi conştienţi artificiali. Există o abundenţă de agenţi conştienţi naturali, suficienţi să administreze orice sarcină rezervată acestor entităţi speciale şi privilegiate. Avem nevoie de instrumente inteligente. Instrumentele nu au drepturi şi nu ar trebui să aibă sentimente care pot fi rănite sau să poată să răspundă cu resentimente la „abuzurile" la care sunt supuse de către utilizatorii inepţi[22]. Unul dintre motivele pentru care nu ar trebui să facem agenţi conştienţi artificiali este că, oricât de autonomi ar deveni (şi, în

principiu, pot deveni la fel de autonomi, de perfectibili şi de creativi ca orice persoană), aceştia nu ar împărtăşi cu noi – fiură instrucţiuni speciale, care pot fi anulate –, agenţii conştienţi naturali, vulnerabilităţile şi mortalitatea noastră.

I-am provocat la un moment dat pe studenţii de la un seminar de la Tufts pe care l-am ţinut împreună cu Matthias Scheutz pe tema agenţilor artificiali şi autonomiei: daţi-mi specificaţiile unui robot care ar putea semna un contract cu putere juridică pentru voi – nu ca surogat pentru un proprietar uman, ci pe cont propriu. Problema nu este să-l faci să înţeleagă clauzele sau să-l faci să mişte un stilou pe o bucată de hârtie, ci de a avea şi de a *merita* statutul juridic de agent moral responsabil. Copiii nu pot semna astfel de contracte, şi nici persoanele cu dizabilităţi al căror statut juridic le cere să fie sub îngrijirea sau responsabilitatea unui tutore sau a altui fel de responsabil. Problema pentru roboţii care ar vrea să obţină un astfel de statut solemn este că, la fel ca Superman, sunt prea invulnerabili pentru a putea face o promisiune credibilă. Dacă nu şi-ar respecta obligaţiile asumate, ce s-ar întâmpla? Care ar fi pedeapsa pentru încălcarea promisiunilor? Să fie închişi într-o celulă sau, mai probabil, să fie demontaţi? Să fie închisă nu ar fi o mare problemă pentru o inteligenţă artificială decât dacă instalăm în prealabil plăcerea de a călători care nu poate fi ignorată sau dezactivată de inteligenţa artificială pe cont propriu (şi ar fi sistematic foarte dificil să implementăm o astfel de soluţie cu succes dată fiind isteţimea şi cunoaşterea de sine presupuse ale inteligenţei artificiale); iar demontarea unui sistem de inteligenţă artificială (fie el robot sau agent imobilizat la pat cum este Watson) nu înseamnă să-l omori dacă informaţiile stocate în designul şi în softul sale sunt păstrate. Uşurinţa înregistrărilor şi transmiterilor digitale – inovaţia care le permite soiturilor şi datelor să fie practic nemuritoare – scoate roboţii din lumea vulnerabilului (cel puţin roboţii pe care ni-i putem imagina de obicei, cu softuri şi memorii digitale). Dacă nu este deja evident, gândeşte-te cum ar fi afectată moralitatea omenească dacă am putea să facem „copii de rezervă” pentru oameni în fiecare săptămână, de exemplu. Te arunci sâmbătă cu capul în jos de pe un pod înalt fără coardă şi ai un iureş de trăiri pe care nu ţi-l vei aduce aminte duminică dimineaţă, când se activează copia de rezervă făcută vineri seară, dar te poţi bucura de clipul video al morţii tale.

Aşadar, ceea ce creăm nu este sau nu ar trebui să fie agenţi conştienţi, umanoizi, ci o entitate cu totul nouă, asemenea oracolelor, fără conştiinţă, fără teamă de moarte, fără iubire şi ură care să o distragă, fără personalitate (dar cu tot soiul de metehne şi ciudăţenii care vor fi identificate fără îndoială drept „personalitatea” sistemului): cutii de adevăruri (dacă suntem norocoşi) aproape sigur contaminate

cu o difuzie de falsuri. Va fi suficient de greu să învăţăm să trăim cu ele, fără să ne mai torturăm cu fantezii despre singularitate în care aceste inteligenţe artificiale ne vor subjuga, la propriu. Folosirea omului *de către om* va fi schimbată în curând – încă o dată – pentru totdeauna, dar putem să prindem bine cârma şi să navigăm printre câteva dintre pericole dacă ne asumăm responsabilitatea traiectoriei.

Capitolul 6

Debandada inumană în care ne-au băgat maşinile

Rodney Brooks

Rodney Brooks este informatician, profesor Panasonic de robotică, emerit, MIT, fostul administrator al Laboratorului de Inteligenţă artificială de la MIT şi al Laboratorului de Informatică şi Inteligenţă Artificială de la MIT Este autorul cărţii Flesh and Machines.

Roboticianul Rodney Brooks, care a apărut în documentul lui Errol Morris din 1997 intitulat Fast, Cheap and Out of Control alături de un dresor de lei, un horticultor şi un expert în şobolanii-cârtiţă golaşi, a fost descris de un cronicar ca fiind „zâmbitor cu o sclipire de nebunie în ochi". Dar acest lucru se poate spune despre aproape toţi vizionarii.

Câţiva ani mai târziu în cariera sa, Brooks, aşa cum se cade unuia dintre cei mai mari roboticieni din lume, a sugerat că „supraantropomorfizăm oamenii, care sunt, la urma urmei, nişte simple maşini". A continuat prin prezentarea unei viziuni plăcute a unei lumi viitoare a inteligenţelor artificiale în care „distincţia dintre noi şi roboţi va dispărea". A recunoscut, de asemenea, şi că are o perspectivă împărţită asupra lumii. „Asemenea unui cercetător religios, păstrez două seturi de credinţe incongruente şi acţionez asupra fiecăruia în diverse circumstanţe", scria el. „Consider că această transcendenţă între sistemele de credinţe va permite omenirii să accepte în cele din urmă roboţii ca maşini emoţionale şi apoi să înceapă să empatizeze cu ei şi să le confere liber-arbitru, respect şi, în cele din urmă, drepturi".

Asta se întâmpla în 2002. În aceste pagini, abordează o perspectivă mai caustică, fie ea mai restrânsă; este alarmat de măsura în care am ajuns să ne bazăm pe sistemele generalizate, care nu sunt numai exploatatoare, ci şi vulnerabile, ca urmare a dezvoltării mult prea rapide a ingineriei software – progres ce pare că a întrecut impunerea siguranţelor eficiente şi fiabile.

Matematicienii şi oamenii de ştiinţă sunt adesea limitaţi în modul în care percep imaginea de ansamblu, dincolo de domeniul lor de

activitate, din cauza instrumentelor şi metaforelor pe care le folosesc în munca lor. Norbert Wiener nu este o excepţie, şi cred că nici eu.

Când a scris *Folosirea omului de către om* Wiener era la sfârşitul erei înţelegerii maşinilor şi animalelor ca simple procese fizice şi la începutul erei noastre curente de înţelegere a maşinilor şi animalelor ca procese computaţionale. Bănuiesc că vor fi ere viitoare ale căror instrumente vor fi la fel de diferite de instrumentele celor două ere la intersecţia cărora s-a aflat Wiener pe cât de diferite sunt instrumentele celor două ere între ele.

Wiener a fost un titan al erei anterioare şi s-a folosit de instrumentele inventate încă de pe vremea lui Newton şi Leibniz pentru a descrie şi a analiza procesele continue din lumea fizică. În 1948, a publicat *Cibernetica*, un cuvânt pe care l-a născocit pentru a descrie ştiinţa comunicării şi controlului atât la maşini, cât şi la animale. Astăzi ne referim la ideile din această carte drept teoria controlului, o disciplină indispensabilă pentru proiectarea şi analiza maşinilor fizice, neglijând în mare parte spusele lui Wiener despre ştiinţa comunicării. Inovaţiile lui Wiener au avut la bază în mare parte munca sa din timpul celui de-al Doilea Război Mondial cu mecanisme de ţintire şi arme de foc antiaeriene. A adus rigoare matematică în proiectarea tehnologiei ale cărei procese de proiectare erau în mare parte euristice: de la sistemele de alimentare cu apă ale romanilor până la motorul cu aburi al lui Watt şi dezvoltarea primelor automobile.

Ne putem imagina o versiune contingenţă diferită a istoriei noastre intelectuale şi tehnologice dacă nu ar fi apărut în scenă Alan Turing şi John von Neumann, ambii aducând contribuţii majore fundaţiei proceselor de calcul. Turing a contribuit cu un model de calcul fundamental – cunoscut acum ca maşina Turing – în lucrarea sa *Despre numerele computabile cu aplicare pe problema deciziei*[23], scrisă şi revizuită în 1936 şi publicată în 1937. În aceste maşini, o bandă liniară de simboluri dintr-un alfabet finit codează intrarea pentru o problemă de calcul, punând la dispoziţie şi mediul de lucru pentru calcul. Era necesară o maşină diferită pentru fiecare problemă de calcul; lucrările ulterioare scrise de alţi cercetători vor arăta că pe una dintre maşini, cunoscută drept Maşina Turing Universală, un set arbitrar de instrucţiuni de calcul poate fi codat pe aceeaşi bandă.

În anii 1940, von Neumann a dezvoltat o maşină abstractă cu autoreproducere denumită automatul celular. În acest caz, ocupa un subset finit dintr-o matrice bidimensională infinită de pătrate, fiecare conţinând un singur simbol dintr-un alfabet finit cu douăzeci şi nouă de simboluri diferite – restul matricei infinite începe goală. Simbolurile individuale din fiecare pătrat se schimbă la unison în baza unei reguli complexe, dar finite, aplicată simbolului curent din pătrat

şi vecinilor săi imediaţi. Sub regula complexă pe care a dezvoltat-o von Neumann, majoritatea simbolurilor din majoritatea pătratelor rămân neschimbate şi câteva se schimbă în fiecare fază. Prin urmare, dacă cineva priveşte pătratele care nu sunt goale, pare că este vorba despre o structură constantă în care se desfăşoară o mică activitate. Când maşina abstractă a lui von Neumann se reproducea, făcea o copie a sa în altă regiune a planului. În cadrul „maşinii" se afla o linie orizontală de pătrate care acţiona ca o bandă liniară finită, utilizând un subset din alfabetul finit. Simbolurile din acele pătrate codau maşina din care făceau parte. În timpul reproducerii maşinii, „banda" se putea deplasa la stânga sau la dreapta şi era interpretată (transcrisă) ca instrucţiuni (conversie) pentru noua „maşină" care era construită, apoi copiată (replicată) –, iar noua copie era pusă în interiorul noii maşini pentru reproducere ulterioară. Francis Crick şi James Watson au arătat mai târziu, în 1953, cum o astfel de bandă poate fi particularizată în biologie de o moleculă lungă de ADN[24] cu alfabetul său finit format din patru baze azotate: guanină, citozină, adenină şi timină (G, C, A şi T)[25] La fel ca în cazul maşinii lui von Neumann, în reproducerea biologică, secvenţa liniară de simboluri din ADN este interpretată – prin transcriere în molecule AM, care sunt apoi convertite în proteine, structurile care creează o celulă nouă – şi apoi ADN-ul este replicat şi încapsulat în noua celulă.

A doua lucrare fundamentală a fost în 1945, într-un raport „versiune preliminară" privind designul unui computer digital, unde Neumann susţinea o memorie care putea conţine atât instrucţiuni, cât şi date. Acum este cunoscut sub numele de computer cu arhitectură Neumann – diferit de computerele cu arhitectură Harvard în care există două memorii separate, una pentru instrucţiuni şi una pentru date. Vasta majoritate de procesoare pentru computere construite în epoca Legii lui Moore sunt bazate pe arhitectura von Neumann, inclusiv cele care stau la baza centrelor noastre de date, laptopurilor şi telefoanelor noastre inteligente. Arhitectura de computer digital a lui von Neumann este conceptual aceeaşi generalizare – de la computerele digitale timpurii construite cu relee electromagnetice la Universitatea Harvard şi Bletchley Park – care apare când treci de la maşină

Turing specială la o maşină Turing universală. În plus, automatele sale cu replicare automată împărtăşesc o similitudine fundamentală atât cu construcţia unei maşini Turing, cât şi cu mecanismele celulelor biologice care se reproduc pe bază de ADN. Chiar şi în ziua de astăzi mai există dezbateri între intelectuali dacă von Neumann a văzut interconexiunile dintre aceste trei lucrări, a lui Turing şi cele două ale sale. Revizuirea făcută de Turing lucrării sale a fost făcută când el şi von Neuman erau ambii la Princeton; într-adevăr, după ce şi-a obţinut

doctoratul, Turing aproape că a rămas în calitate de postdoctorand al lui Neumann.

Fără Turing şi von Neumann, cibernetica lui Wiener ar fi putut rămâne un mod dominant de gândire şi de structurare a tehnologiei mult mai mult timp decât scurta sa perioadă de supremaţie. În această versiune imaginară de istorie, am fi putut trăi astăzi într-o lume de fapt *steam-punky* nu doar să observăm particularizările sale fantastice la Maker Faires!

Ce vreau să spun este că Wiener privea lumea – fizică, biologică şi (în *sens uman*) sociologică – într-un mod foarte special. A analizat lumea ca un set continuu de variabile, după cum explică în capitolul 1 împreună cu o apreciere a termodinamicii prin lentila statisticilor lui Gibbs. De asemenea, încearcă să introducă forţat un model slab şi neconvingător al informaţiei ca un mesaj care se transmite între entităţi fizice şi biologice. Pentru mine şi din punctul de vedere al informaţiilor de astăzi, după şaptezeci de ani, instrumentele sale par cumplit de inadecvate pentru descrierea mecanismelor care stau la baza sistemelor biologice, iar astfel nu a înţeles cum mecanisme similare ar putea până la urmă să fie încorporate în sisteme de calcul tehnologice – aşa cum sunt încorporate acum. Tehnologiile dominante de astăzi au fost dezvoltate în lumea lui Turing şi von Neumann, mai degrabă decât în lumea lui Wiener, în prima revoluţie industrială, energia de la un motor cu abur sau de o roată de moară a fost utilizată de muncitorii umani pentru a înlocui propria lor energie. În loc să fie o sursă de energie pentru munca fizică, oamenii au devenit modulatorii acum va fi utilizată o sursă puternică de energie. Dar, din cauza faptului că motoarele cu abur şi roţile de moară trebuiau să fie mari pentru a fi o utilizare rentabilă a capitalului şi pentru că în secolul al XVIII-lea singura tehnologie pentru distribuirea spaţială a energiei era mecanică şi funcţiona doar pe o rază foarte mică, mulţi muncitori erau nevoiţi să se înghesuie în jurul sursei de energie. Wiener argumentează corect că abilitatea de a transmite energie ca electricitate a determinat o a doua revoluţie industrială. Acum sursa de energie poate fi la o distanţă mare de locul de utilizare şi, de la începutul secolului XX, producţia a putut fi mult mai dispersată după construirea reţelelor de distribuţie a energiei electrice.

Wiener argumentează mai departe că încă o tehnologie nouă, cea maşinilor de calcul care luau naştere în vremea sa, va da naştere încă unei revoluţii. Maşinile despre care vorbeşte par să fie atât analoge, cât şi (probabil) digitale; şi evidenţiază, în *The Human Use of Human Beings*, faptul că, de vreme ce vor putea lua decizii, atât muncitorii din producţie, cât şi cei de la birouri ar putea fi reduşi la a fi simple rotiţe într-o maşinărie mult mai mare. Se temea că oamenii s-ar putea folosi şi abuza unii pe alţii prin structuri organizaţionale pe care această

capacitate le va încuraja. Cu siguranţă, am trăit acest scenariu în ultimii şaizeci de ani şi această perturbare nici măcar nu se apropie de final.

Însă perspectiva sa bazată pe fizică în ceea ce priveşte calculul îl proteja de realizarea a cât de rău ar putea sta lucrurile cu adevărat. Vedea capacitatea maşinilor de a comunica drept furnizarea unei modalităţi noi şi mai inumane de a exercita comanda şi controlul. Nu a întrevăzut faptul că, în câteva decenii, sistemele de calcul vor deveni mai asemănătoare sistemelor biologice şi se pare, din descrierile sale din capitolul 10 din propria sa lucrare despre modelarea anumitor aspecte ale biologiei, că a subapreciat grav multele ordine de magnitudine de complexitate ulterioare ale biologiei peste cele ale fizicii. Astăzi suntem într-o situaţie mult mai complexă decât a prevăzut el şi mă tem că este mult mai periculos decât în cele mai sumbre predicţii ale sale.

În anii 1960, calculul a început să se bazeze ferm pe fundaţia stabilită de Turing şi de von Neumann şi aceasta era reprezentată de calculul digital, bazat pe ideea unor alfabete finite pe care le-au utilizat ambii. O secvenţă de o lungime arbitrară sau un *seringa* formată din caracterele unui alfabet finit poate fi codată ca un număr întreg unic. Aşa cum se întâmplă şi cu maşinile Turing, formalismul calculului a devenit formalismul calculării unei funcţii cu valori întregi pentru o singură intrare evaluată ca număr întreg.

Turing şi von Neumann au murit ambii în anii 1950 şi, la momentul respectiv, aşa vedeau ei calculul. Niciunul dintre ei nu a prevăzut creşterea exponenţială a puterii de calcul pe care urma să o aducă Legea lui Moore – şi nici cât de universale urmau să devină maşinile de calcul. Nu a prevăzut nici dezvoltarea a două modele pentru sistemele noastre de calcul, fiecare dintre ele prezentând o mare ameninţare pentru societatea umană.

Primul îşi are rădăcina în abstracţiile adoptate. În cei cincizeci de ani de cursă alimentată de Legea lui Moore de a produce software care să exploateze dublarea capacităţii computerelor la fiecare doi ani, grija şi certificarea disciplinelor de inginerie tipice au încetat să mai fie folosite. Dezvoltarea soiturilor era rapidă şi supusă erorilor. Această dezvoltare rapidă a soiturilor fără standarde de corectitudine a deschis multe rute de exploatare a arhitecturii Neumann de stocare a datelor şi instrucţiunilor în aceeaşi memorie. Una dintre cele mai comune rute, cunoscută sub numele de „depăşire a memoriei *buffer*”[26] implică un număr de intrare (sau un şir lung de caractere) care este mai mare decât a prevăzut programatorul şi depăşeşte memoria revărsându-se în zona în care sunt stocate instrucţiunile. Dacă proiectează cu atenţie un număr de intrare care este mult prea mare, cineva care foloseşte un soft îl poate infecta cu instrucţiuni care nu

sunt prevăzute de programator, modificând astfel ceea ce face. Acestea sunt bazele creării unui virus de computer – numit astfel datorită similitudinii sale cu un virus biologic. Cel din urmă injectează ADN suplimentar într-o celulă, iar mecanismul de transcriere şi traducere al celulei respective îl interpretează orbeşte făcând proteine care pot fi dăunătoare celulei-gazdă. În plus, mecanismul de replicare al celulei are grijă de multiplicarea virusului. Astfel o entitate străină mică poate prelua controlul asupra unei entităţi mult mai mari şi-i poate modifica comportamentul în moduri neprevăzute.

Acestea şi alte forme de atacuri digitale ne-au privat de securitatea vieţilor noastre de zi cu zi. Ne bazăm pe computere pentru aproape orice acum. Ne bazăm pe computere pentru infrastructura noastră de electricitate, gaze, drumuri, maşini, trenuri şi avioane; acestea sunt toate vulnerabile. Ne bazăm pe computere pentru *banking*} plata facturilor, conturile noastre de economii, ipotecile, achiziţiile de bunuri şi servicii – şi acestea sunt toate vulnerabile. Ne bazăm pe computere pentru distracţie, comunicaţii personale şi profesionale, securitatea noastră fizică acasă, informaţiile noastre despre lume şi sistemele noastre de vot – toate vulnerabile. Niciuna dintre acestea nu va fi reparată prea curând. Între timp, multe aspecte ale societăţii noastre sunt deschise unor atacuri violente, fie venite din partea unor infractori liber-profesionişti, fie din partea adversarilor statului.

A doua dezvoltare este că calculul a trecut de simplele funcţii de calcul. În schimb, programele rămân online continuu, putând aduna astfel date despre o secvenţă de interogări. În schema Wiener / Turing / von Neumann, putem privi tiparul comunicării pentru un browser web drept:

Utilizator: dă-mi pagina web A.
Browser: iată pagina web A.

Utilizator: dă-mi pagina web B.
Browser: iată pagina web B.

Acum, în schimb, poate arăta aşa:

Utilizator: dă-mi pagina web A.
Browser: iată pagina web A. (Şi voi reţine în secret că mi-ai cerut pagina web A.)

Utilizator: dă-mi pagina web B.
Browser: iată pagina web B. Văd o corelaţie între conţinutul ei şi cel al cererii de mai devreme, pagina web A, aşa că-mi voi actualiza modelul pe care îl am pentru tine, utilizatorul, şi-l voi transmite companiei care m-a produs.

Când maşina nu mai calculează pur şi simplu o funcţie, ci întreţine în schimb o stare, poate începe să facă inferenţe despre om în funcţie de secvenţa de cereri care i-a fost prezentată. Şi când diverse programe corelează mai multe fluxuri de cereri – de exemplu, corelează căutările de pagini web cu postările social-media sau plata serviciilor pe altă platformă sau timpul pe care l-ai petrecut vizionând o reclamă sau unde a fost condus utilizatorul cu GPS-ul telefonului inteligent –, sistemele totale ale mai multor programe ce comunică între ele şi cu diverse baze de date conduc la o cu totul nouă pierdere a confidenţialităţii. Marele salt exploatator făcut de atât de multe companii de pe Coasta de Vest a fost să monetizeze aceste inferenţe fără permisiunea conştientă a persoanei care generează interacţiunile cu platformele maşinilor de calcul.

Wiener, Turing şi von Neumann nu au putut prevedea complexitatea acestor platforme, în timp ce bălmăjeala în termeni juridici din contractele condiţiilor de utilizare pe care oamenii la acceptă voit, fără să aibă cea mai mică idee ce presupun acestea, îi face să renunţe la drepturi la care altfel nici nu s-ar gândi să renunţe într-o interacţiune unu la unu cu o persoană. Platformele de calcul au devenit un scut în spatele căruia anumite companii se ascund pentru a-i exploata inuman pe alţii. În alte ţări, guvernele sunt cele care fac aceste manipulări şi acolo obiectivul nu este profitul, ci suprimarea disidenţei.

Omenirea s-a pus într-o situaţie dificilă: suntem exploataţi de companii care, paradoxal, ne oferă servicii pe care ni le dorim şi, în acelaşi timp, vieţile noastre depind de multe sisteme bazate pe soituri care sunt deschise atacurilor. Să ieşim din această încurcătură este un proiect pe termen lung. Va necesita planificare, legi şi, cel mai important, *leadersbip* moral. *Leadership*-ul moral este prima şi cea mai mare provocare.

Capitolul 7
Unitatea inteligenţei

Frank Wilczek

Frank Wilczek este profesor Herman Feshbach de fizică la MIT, laureat al Premiului Nobel pentru fizică în 2004 şi autorul cărţii ABeautiful Question: Finding Natures Deep Design.

L-am cunoscut pe Frank Wilczek în anii 1980, când m-a invitat acasă la el, în Princeton, pentru a discuta despre anioni. „Adresa este strada Mercer, nr. 112", mi-a scris el. „Caută o casă fără străduţă de parcare". Şi iată-mă după câteva ore în vechea

cameră de zi a lui Einstein, discutând cu un viitor laureat al Premiului Nobel în fizică. Dacă Frank era la fel de impresionat ca mine de împrejurimi, nu o să ghiciţi niciodată. Singurul său comentariu era despre dificultatea de a găsi un loc de parcare în faţa unei „case fără străduţă de parcare".

Spre deosebire de majoritatea fizicienilor teoreticieni, Frank era interesat de mult timp de inteligenţa artificială, mărturie stând aceste trei „observaţii":

1. *„Francis Crick i-a spus „ipoteza uimitoare": şi anume, conştiinţa, cunoscută şi ca minte, este o proprietate emergentă a materiei", afirmaţie care, dacă este adevărată, arată că „toată inteligenţa este inteligenţă automată. Ce diferenţiază inteligenţa naturală de cea artificială nu este natura ei, ci modul în care este creată".*
2. *„Inteligenţa artificială nu este produsul unei invazii a extratereştrilor. Este un artefact al unei culturi omeneşti speciale şi reflectă valorile acelei culturi".*
3. *„Afirmaţia şocantă a lui David Hume «raţiunea este şi trebuie să fie sclava pasiunilor» a fost scrisă în 1783 şi trebuia să se aplice, bineînţeles, numai raţiunii şi pasiunilor omeneşti… Dar ideea logică/filosofică a lui Hume rămâne valabilă şi pentru inteligenţa artificială. Mai simplu: stimulentele, şi nu logica abstractă, stau la baza comportamentului".*

Observă că „marea poveste a secolelor 20 şi 21 este că, odată cu dezvoltarea computerelor, învăţăm să calculăm consecinţele legilor fundamentale din ce în ce mai bine. Există şi un ciclu de *feedback*. când înţelegi materia mai bine, poţi proiecta computere mai bune, care-ţi vor permite să calculezi mai bine. Este un fel de spirală ascendentă" Aici, argumentează că inteligenţa umană deţine avantajul, deocamdată – însă viitorul nostru, nelegat de sistemul nostru solar şi fără îndoială nici de galaxia noastră, nu se va înfăptui niciodată fără ajutorul inteligenţelor noastre artificiale.

I. Un răspuns simplu la întrebări contencioase

** Poate o inteligenţă artificială să fie conştientă?*
** Poate o inteligenţă artificială să fie creativă?*
** Poate o inteligenţă artificială să fie malefică?*

Aceste întrebări sunt puse adesea astăzi, atât în mass-media populară, cât şi în dezbaterile ştiinţifice informate. Dar discuţiile nu par să fie niciodată convergente. Aici voi începe prin a oferi

răspunsuri după cum urmează:

În baza psihologiei fiziologice, neurobiologiei şi fizicii, ar fi surprinzător dacă răspunsurile nu ar şi de da şi da. Motivul este simplu şi totuşi profund: dovezile din domeniile respective fac foarte posibil faptul că nu există nicio împărţire clară între inteligenţa naturală şi cea artificială.

În cartea sa din 1994 pe această temă, renumitul biolog Francis Crick a propus o „ipoteză uluitoare": că mintea ia naştere din materie. A spus că mintea, în toate aspectele sale, „nu este nimic mai mult decât un ansamblu de celule nervoase şi moleculele lor asociate".

„Ipoteza uluitoare" este de fapt fundamentarea neuroştiinţei moderne. Oamenii încearcă să înţeleagă cum funcţionează minţile înţelegând cum funcţionează creierul şi încearcă să înţeleagă cum funcţionează creierul studiind cum este codată informaţia în semnale electrice şi chimice, transformate de procese fizice şi cum este utilizată pentru a controla comportamentul. În această încercare ştiinţifică, nu permit comportamente extrafizice. Până acum, în mii de experimente minunate, această strategie nu a dat greş niciodată. Nu s-a dovedit niciodată necesar să permitem influenţei conştiinţei sau creativităţii eliberate de activitatea creierului să explice un eveniment psihofizic sau neurobiologic observat. Nimeni nu a întâlnit vreodată o putere a minţii care să fie separată de evenimente fizice convenţionale din organismele biologice. Deşi sunt multe lucruri pe care nu le înţelegem cu privire la creier şi minte, „ipoteza uluitoare" a rămas intactă.

Dacă ne lărgim perspectiva dincolo de neurobiologie şi luăm în considerare întregul spectru de experimente ştiinţifice, cazul devine şi mai fascinant. În fizica modernă, centrele de interes sunt adesea fenomene foarte delicate. Pentru a le investiga, experimentatorii trebuie să ia multe măsuri de precauţie împotriva contaminării cu „zgomot". Adesea li se pare necesar să construiască protecţie suplimentară contra electricităţii statice şi câmpurilor magnetice; să compenseze pentru vibraţii mărunte datorate microcutremurelor sau maşinilor în trecere; să funcţioneze la temperaturi extrem de mici şi în vid cu presiune; şi aşa mai departe. Dar există o excepţie notabilă: nu li s-a părut niciodată necesar să acorde permisiuni pentru ce gândesc oamenii din apropiere (sau din depărtare). Nu par să existe „unde mentale", diferite de procesele fizice cunoscute, capabile să influenţeze evenimentele fizice.

Acea concluzie, luată la valoare nominală, şterge distincţia dintre inteligenţa naturală şi artificială. Implică faptul că, dacă ar fi să duplicăm sau să simulăm cu precizie procesele fizice care apar în creier – aşa cum, în principiu, putem – şi să conectăm intrările şi ieşirile la organe de simţ şi la muşchi, atunci am reproduce, într-un artefact fizic, manifestările observate ale inteligenţei naturale. Nu ar

lipsi nimic observabil. Ca observator, nu aş avea motive în plus sau în minus să presupun conştiinţă, creativitate sau răutate acelui artefact pentru care nu atribui acele proprietăţi omologilor săi naturali, cum ar fi alte fiinţe umane.

Astfel, combinând „ipoteza uluitoare" a lui Crick din neurobiologie cu dovezi puternice din fizică, deducem că inteligenţa naturală este un caz special de inteligenţă artificială. Această concluzie merită un nume şi o voi numi „corolarul uluitor".

Cu ea avem răspunsul la cele trei întrebări ale noastre. De vreme ce conştiinţa, creativitatea şi răutatea sunt trăsături evidente ale inteligenţei naturale omeneşti, sunt şi trăsături posibile ale inteligenţei artificiale.

Cu o sută de ani în urmă sau chiar cincizeci, să crezi că mintea ia naştere din materie şi să inferezi corolarul nostru că inteligenţa naturală este un caz special de inteligenţă artificială, ar fi presupus un salt de credinţă. În lumina multelor lipsuri conexe – abisuri, de fapt – din înţelegerea contemporană a biologiei şi fizicii, acestea erau propuneri cu adevărat îndoielnice. Dar dezvoltările epocale din acele domenii au schimbat imaginea:

În biologie: cu un secol în urmă, nu numai gândurile, ci şi metabolismul, ereditatea şi percepţia erau aspecte profund misterioase ale vieţii care sfidau explicaţiile fizice. Astăzi, bineînţeles, avem relatări extrem de bogate şi detaliate despre metabolism, ereditate şi multe aspecte ale percepţiei, de jos în sus, începând de la nivelul molecular.

În fizică: după un secol de fizică cuantică şi aplicarea ei pe materiale, fizicienii au descoperit, în repetate rânduri, cât de bogat şi ciudat poate fi comportamentul materiei. Superconductorii, laserele şi multe alte minuni demonstrează că ansamblele mari de unităţi moleculare, fiecare simplu în sine, pot prezenta comportamente „emergente", calitativ noi, rămânând complet obediente legilor fizicii. Chimia, inclusiv biochimia, este cornul abundenţei în ceea ce priveşte fenomenele emergente, toate fiind acum înrădăcinate puternic în fizică. Fizicianul deschizător de drumuri Philip Anderson, într-un eseu intitulat „Mai mult este diferit", prezintă o discuţie clasică despre emergenţă. Începe prin a recunoaşte că „ipoteza reducţionistă [adică, integralitatea explicaţiilor fizice bazată pe interacţiunile cunoscute ale părţilor simple] poate fi o simplă controversă în rândurile filosofilor, dar consider că este acceptată fără echivoc de vasta majoritate a oamenilor de ştiinţă activi". Însă continuă prin a evidenţia că „comportamentul agregatelor mari şi complexe de particule elementare, se pare, nu trebuie înţeles în termenii unei simple extrapolări a proprietăţilor câtorva particule"[27]. Fiecare nivel nou de mărime şi complexitate suportă forme noi de organizare, ale căror

tipare codează informaţii în moduri noi şi al căror comportament este cel mai bine descris în concepte noi.

Computerele electronice sunt exemple magnifice de emergenţă. Aici toate cărţile sunt pe faţă. Inginerii proiectează cu regularitate, de jos în sus, în baza unor principii fizice cunoscute (şi destul de sofisticate), maşini care prelucrează informaţiile în moduri extrem de impresionante. iPhone-ul tău te bate la şah, adună şi îţi pune la dispoziţie rapid informaţii despre orice şi mai face şi nişte poze minunate. Deoarece procesul prin care computerele, telefoanele inteligente şi alte obiecte inteligente sunt produse este complet transparent, nu încape îndoială despre capacităţile lor minunate că iau naştere din procese fizice obişnuite, pe care le putem urmări până la nivelul electronilor, fotonilor, quarcilor şi gluonilor. Evident, materia brută poate deveni destul de inteligentă.

Daţi-mi voie să rezum argumentul. Din două ipoteze cu sprijin puternic, am tras o concluzie simplă:

- *Mintea omenească ia naştere din materie.*
- *Materia este definită de fizică.*
- *Aşadar, mintea umană ia naştere din procese fizice pe care le înţelegem şi pe care le putem reproduce artificial.*
- *Aşadar, inteligenţa naturală este un caz special de inteligenţă artificială.*

Bineînţeles, „corolarul nostru uluitor” poate da greş; primele două linii ale acestui argument sunt ipoteze. Dar eroarea lor ar conduce la o descoperire fundamentală – un fenomen semnificativ nou, cu consecinţe fizice la scară largă, care are loc în circumstanţe fizice neremarcabile şi bine studiate (de exemplu, materialele, temperaturile şi presiunile din creierele omeneşti), însă care a reuşit cumva să se sustragă multe decenii arsenalului sofisticat de instrumente al investigatorilor. O astfel de descoperire ar fi… uluitoare.

II. Viitorul inteligenţei

Face parte din natura umană să-şi îmbunătăţească mintea şi corpul. Istoric, vestimentaţia, ochelarii şi ceasurile sunt exemple de augmentări din ce în ce mai sofisticate care ne îmbunătăţesc rezistenţa, percepţia şi conştiinţa. Sunt îmbunătăţiri majore ale zestrei umane naturale a căror familiaritate nu trebuie să ne orbească în ce priveşte profunzimea lor. Astăzi telefoanele inteligente şi internetul aduc dorinţa umană de augmentare în domenii mai centrale identităţii noastre de fiinţe inteligente. Ne conferă, în consecinţă, acces rapid la o conştiinţă colectivă vastă şi la memorie colectivă vastă.

În acelaşi timp, inteligenţe artificiale autonome au devenit

campioni mondiali într-o gamă variată de jocuri „cerebrale”, cum ar fi şahul şi Go, şi au preluat controlul asupra sarcinilor sofisticate de recunoaştere a tiparelor, cum ar fi reconstituirea întâmplărilor din timpul reacţiilor complexe ce au loc în Large Hadron Collider, de la furtuna de particule emergente până la găsirea unor particule noi; sau colectarea indiciilor dintr-o radiografie neclară, un RMN funcţional şi alte tipuri de imagini pentru diagnosticarea problemelor medicale.

Unde ne duce această dorinţă de autoîmbunătăţire şi inovaţie? Deşi secvenţa exactă de evenimente şi scara de timp pe care se desfăşoară sunt imposibil de preconizat (sau, cel puţin, pe mine mă depăşesc), câteva idei de bază sugerează că, în cele din urmă, cele mai puternice întrupări ale minţii vor fi foarte diferite de creierele umane aşa cum le cunoaştem noi astăzi.

Să ne gândim la şase factori prin care tehnologiile de procesare a informaţiei depăşesc capacităţile umane – vast, calitativ, sau ambele:

- Viteză: *mişcarea orchestrată a electronilor, care este inima procesării moderne a informaţiei artificiale, poate fi mult mai rapidă decât procesele de difuzie şi schimbare chimică prin care funcţionează creierele. Frecvenţele procesoarelor moderne se apropie de 10 GHZ, ceea ce înseamnă 10 miliarde de operaţii pe secundă. Nicio măsură a vitezei nu se aplică varietăţii năucitoare a proceselor creierului, dar o limitare fundamentală este latenţa potenţialelor de acţiune, care le limitează spaţierea la câteva zeci pe secundă. Probabil nu este niciun accident că „frecvenţa cadrelor” la care putem distinge că filmele sunt de fapt secvenţe de imagini este de aproximativ 40 pe secundă. Prin urmare, procesarea electronică este de aproape un miliard de ori mai rapidă.*
- Dimensiune: *dimensiunea liniară a unui neuron tipic este de aproximativ 10 microni. Dimensiunile moleculare, care stabilesc o limită practică, sunt de aproximativ 10.000 de ori mai mici şi unităţile de procesare artificiale se apropie de această scară. Dimensiunile reduse fac comunicarea mai eficientă.*
- Stabilitate: *în timp ce memoria umană este esenţial continuă (analogă), memoria artificială poate încorpora caracteristici (digitale) discrete. În timp ce cantităţile analoge se pot eroda, cantităţile digitale pot fi stocate, împrospătate şi întreţinute cu o acurateţe completă.*
- Ciclul de lucru: *creierele omeneşti obosesc de la efort. Au nevoie de pauze pentru a se hrăni şi pentru a dormi. Poartă povara îmbătrânirii. Şi, cel mai important: mor.*

- Modularitate *(arhitectură deschisă): deoarece procesoarele de inteligenţă artificială pot accepta interfeţe digitale precis definite, pot asimila cu uşurinţă modele noi. Astfel, dacă vrem ca un computer să „vadă" ultraviolete sau infraroşii sau să „audă" ultrasunete, putem alimenta ieşire direct din senzorul corespunzător în „sistemul său nervos". Arhitectura creierelor este mult mai închisă şi opacă, iar sistemul imunitar uman rezistă activ implanturilor.*
- Promptitudine cuantică*: un caz de modularitate merită o menţiune specială datorită potenţialului său pe termen lung. Recent, fizicienii şi specialiştii în informaţii au ajuns să aprecieze faptul că principiile mecanicii cuantice suportă noi principii de calcul, care pot crea forme calitative noi de prelucrare a informaţiei şi (plauzibil) niveluri noi de inteligenţă. Dar aceste posibilităţi se bazează pe aspectele comportamentului cuantic care sunt foarte delicate şi par să fie necorespunzătoare pentru o interfaţă cu mediul cald, umed şi haotic al creierelor umane.*

Evident, ca platforme pentru inteligenţă, creierele umane sunt departe de a fi optime. Totuşi, deşi roboţii de curăţenie versatili sau soldaţii mecanici îşi găsesc pieţe rentabile, în prezent nu există nicio maşină care să se apropie de tipul de inteligenţă umană cu scopuri generale de care ar avea nevoie acele aplicaţii. În ciuda slăbiciunii lor relative pe multe fronturi, creierele omeneşti au avantaje uriaşe faţă de competiţia lor artificială. Daţi-mi voie să menţionez cinci:

- Tridimensionalitatea: *deşi, aşa cum am observat deja, dimensiunile liniare ale unităţilor de procesare artificiale existente sunt mult mai mici decât cele ale creierelor, procedura prin care sunt create - centrată pe litografie (în principiu, schiţare) - este esenţialmente bidimensională. Acest lucru se vede clar în geometria plăcilor şi cipurilor de computer. Bineînţeles, se pot conecta mai multe plăci, dar spaţierea dintre straturi este mult mai mare şi comunicarea mult mai puţin eficientă decât în cadrul aceluiaşi strat. Creierele folosesc mai bine cele trei dimensiuni.*
- Autoreparaţia: *creierele omeneşti îşi pot reveni sau pot ocoli multe tipuri de vătămări sau erori. Computerele trebuie adesea să fie reparate sau repomite extern.*
- Conectivitatea: *neuronii umani suportă câteva sute de*

conexiuni (sinapse). În plus, tiparul complex al acestor conexiuni este plin de sens. (Vezi următoarea mea idee.) Unităţile computerelor fac doar câteva conexiuni, în tipare regulate şi fixe.

- Dezvoltarea *(asamblare automată cu sculptură interactivă): creierul uman îşi creşte unităţile prin diviziune celulară şi le orchestrează în structuri coerente prin mişcare şi înfăşurare. De asemenea proliferează o abundenţă de conexiuni între celule. O parte importantă a sculptării sale apare prin procese active în timpul copilăriei sale, când individul interacţionează cu mediul său. În acest proces, multe conexiuni sunt eliminate, în timp ce altele sunt fortificate, în funcţie de eficacitatea utilizării. Astfel structura fină a creierului este reglată prin interacţiunea cu lumea externă – o sursă bogată de informaţii şi feedback!*
- Integrarea (senzori şi activatori): *creierul omenesc vine echipat cu o varietate de organe senzoriale, printre care notabili sunt ochii concrescuţi şi cu mecanisme versatile de control, inclusiv mâini care construiesc, picioare care merg şi guri care vorbesc. Aceşti senzori şi mecanisme de control sunt integrate perfect în centrele de procesare a informaţiei din creier, fiind perfecţionate de-a lungul a milioane de ani de selecţie naturală. Interpretăm semnalele lor brute şi controlăm acţiunile lor la scară mare cu o atenţie conştientă minimă. Reversul medaliei este că nu ştim cum facem asta, şi implementarea este opacă. Se dovedeşte a fi surprinzător de dificil să ajungem la standarde umane în ceea ce priveşte aceste funcţii de intrare-ieşire „de rutină”.*

Aceste avantaje ale creierelor omeneşti asupra artefactelor proiectate în prezent sunt profunde. Creierele omeneşti oferă o dovadă de existenţă înălţătoare, arătându-ne câteva metode de a ieşi şi mai mult din materie. Când va ajunge ingineria la acest nivel, dacă va ajunge vreodată?

Nu ştiu sigur, dar daţi-mi voie să vă spun câteva păreri informate. Provocările tridimensionalităţii şi, într-o mai mică măsură, ale autoreparaţiei nu par covârşitoare. Pun nişte probleme inginereşti dificile, dar multe îmbunătăţiri incrementale sunt uşor de imaginat şi se întrevăd căi clare de înaintare. Şi, deşi puterile ochilor omeneşti, ale mâinilor şi ale altor organe senzoriale şi mecanisme de acţionare sunt incredibil de eficiente, capacităţile lor sunt departe de a depăşi orice limite fizice. Sistemele optice pot face poze cu o rezoluţie mai

mare în spaţiu, timp şi culoare şi în mai multe regiuni ale spectrului electromagnetic; roboţii se pot mişca mai repede şi pot fi mai puternici; şi aşa mai departe. În aceste domenii, componentele necesare pentru performanţe supraomeneşti, pe mai multe axe, sunt deja disponibile. Dificultatea rezidă în a introduce şi a scoate rapid informaţia din ele, într-o limbă a unităţilor de prelucrare a informaţiei.

Şi astfel ajungem la restul avantajelor – şi, cred eu, cele mai profunde – creierelor asupra dispozitivelor artificiale, care iau naştere din conectivitatea lor şi din dezvoltarea interactivă. Aceste două avantaje sunt sinergetice, deoarece dezvoltarea interactivă sculptează structura puternic conectată şi vastă a creierului de copil, activată de creşteri exponenţiale de neuroni şi sinapse, pentru a se modela în instrumentul extraordinar care urmează să devină. Informaticienii încep să descopere puterea arhitecturii creierului: reţelele neuronale, al căror design fundamental, aşa cum sugerează şi numele, a fost inspirat direct de designul creierului, au avut succese spectaculoase în domeniul jocurilor şi în recunoaşterea tiparelor, aşa cum s-a observat. Dar ingineria de astăzi nu are nimic comparabil – în domeniul (în prezent) esoteric al maşinilor cu reproducere automată – cu puterea şi versatilitatea neuronilor şi sinapselor lor. Aceasta poate deveni o frontieră nouă şi minunată a cercetării. Şi aici biologia ne poate deschide drumul, pe măsură ce înţelegem dezvoltarea biologică suficient de bine pentru a-i imita esenţa.

În ansamblu, avantajele inteligenţei artificiale faţă de inteligenţa naturală par să fie permanente, iar avantajele inteligenţei naturale faţă de inteligenţa artificială, deşi sunt substanţiale în prezent, par trecătoare. Cred că vor mai trece multe decenii până când ingineria le va prinde din urmă, dar – dacă nu vor apărea războaie catastrofale, schimbări climatice bruşte sau molime, astfel încât progresul tehnologic să poată rămâne viguros – puţine secole.

Dacă estimarea este corectă, putem anticipa câteva generaţii în care oamenii, împuterniciţi şi augmentaţi de dispozitive inteligente, vor coexista cu inteligenţele artificiale autonome din ce în ce mai capabile, fiind astfel supuse unei evoluţii rapide. Date fiind avantajele inerente pe care le vor oferi dispozitivele proiectate în cele din urmă, avangarda acestei evoluţii va fi reprezentată de cyborgi şi superminţi, şi nu de nişte *homo sapiens* ornaţi.

Un alt elan important va veni din explorarea mediilor ostile, atât de pe Pământ (de exemplu, din adâncul oceanelor), cât şi, în special, din spaţiu. Corpul omenesc nu este adaptat la condiţiile care se află în afara unui interval restrâns de temperaturi, presiuni şi compoziţii atmosferice. Are nevoie de o varietate mare de nutrienţi specifici şi complecşi, precum şi de multă apă. De asemenea, nu este obişnuit cu

radiaţiile. După cum a demonstrat clar programul spaţial, este dificil şi scump să ţii oamenii în afara zonei lor de confort terestru. Cyborgii sau inteligenţele artificiale autonome se pot dovedi mult mai eficienţi în aceste explorări. Inteligenţele artificiale cuantice, cu sensibilitatea lor la zgomot, ar putea fi chiar mai fericite în răceala şi întunericul spaţiului interstelar.

Într-un pasaj impresionant din romanul *Odd John* (1935), geniul singular al genului Science fiction, Olaf Stapledon are un erou, o inteligenţă (mutantă) supraomenească, ce descrie specia *homo sapiens* ca „Arheopterixul spiritului". Îi spune acest lucru cu căldură prietenului şi biografului său, care este un om obişnuit. Arheopterix a fost o creatură nobilă şi o punte către creaturi mai măreţe.

Capitolul 8

Să aspirăm la mai mult decât doar să nu devenim depăşiţi

Max Tegmarck

Max Tegmarck este fizician MIT şi cercetător în domeniul inteligenţei artificiale, preşedintele Institutului Future for Life, director ştiinţific al Institutului Foundational Questions şi autorul cărţilor Our Mathematical Universe şi Life 3. O: Being Human in the Age of Artificial Intelligence.

L-am cunoscut pe Max Tegmark cu câţiva ani în urmă, când mi l-a prezentat colegul său de la MIT, Alan Guth, părintele teoriei inflaţiei. Fizician teoretician şi cosmolog distins el însuşi, grija principală a lui Max în prezent este riscul existenţial care stă la pândă implicat de crearea unei inteligenţe artificiale generale (adică, una care rivalizează inteligenţei umane). Cu patru ani în urmă, Max a înfiinţat, împreună cu Jaan Tallinn şi alţii, Institutul Future of Life, care se prezintă ca „o organizaţie deschisă ce lucrează pentru a se asigura că cele mai puternice tehnologii ale viitorului sunt benefice umanităţii". În timpul turneului de prezentare a cărţii sale în Londra, era în focurile planificării pentru Institutul Future of Life şi recunoaşte că a fost mişcat până la lacrimi într-o staţie de metrou după o vizită la Muzeul de Ştiinţe din Londra, cu expoziţii pe tema realizărilor tehnologice ale umanităţii. Tot acest progres impresionant este în zadar? Consiliul ştiinţific al Institutului Future of Life îi include pe Elon Musk, Frank Wilczek, George Church, Stuart Russell şi pe filosoful de la Oxford Nick Bostrom, cel care a conceput un Gedankenexperiment[28] citat adesea, ce are ca urmare o lume plină de agrafe de hârtie şi nimic altceva, produsă de o inteligenţă artificială generală (aparent) binevoitoare, care nu făcea decât să respecte comenzile primite. Institutul sponsorizează conferinţe (Porto Rico 2015, Asilomar 2017) pe tema problemelor de siguranţă create de inteligenţa artificială şi, în 2018, a instituit o competiţie pentru burse ce se concentrează pe cercetarea maximizării beneficiilor sociale ale inteligenţei artificiale generale.

Deşi Max este uneori listat – de necunoscători – în tabăra panicarzilor, el crede, la fel ca şi Frank Wilczek, într-un viitor care va beneficia enorm de inteligenţa artificială generală dacă, în încercarea de a o crea, putem să menţinem pentru aceasta specia umană ca prioritate.

Deşi există o mare controversă despre momentul în care inteligenţa artificială va avea un impact semnificativ asupra

umanităţii, situaţia este mai clară din perspectivă cosmică: viaţa de pe Pământ care dezvoltă tehnologie se grăbeşte să devină depăşită fără să se gândească în mod serios la consecinţe. Acest lucru mi se pare jenant, dat fiind faptul că putem crea oportunităţi uimitoare astfel încât umanitatea să prospere mai mult decât oricând, dacă îndrăznim să-i dăm un curs mai ambiţios.

Universul nostru a devenit conştient de sine după 13, 8 miliarde de ani de la naşterea sa. Pe o planetă mică şi albastră, părţi minuscule de Univers au descoperit că ceea ce au crezut ei cândva că este întreaga existenţă este o parte infimă din ceva mult mai mare: un sistem solar într-o galaxie într-un univers cu peste 100 de miliarde de alte galaxii, aranjate într-un tipar elaborat de grupuri, clustere şi superdustere.

Conştiinţa este trezirea cosmică; a transformat Universul nostru dintr-un zombi fără minte şi fără conştiinţă de sine într-un ecosistem viu ce adăposteşte reflecţie de sine, frumuseţe, speranţă, sens şi scop. Dacă această trezire nu ar fi avut loc niciodată, Universul nostru nu ar fi avut niciun sens – un spaţiu uriaş irosit. Dacă Universul nostru se va stinge permanent din cauza unei calamităţi cosmice sau a unei nenorociri create de noi înşine, va deveni din nou fără sens.

Pe de altă parte, lucrurile s-ar putea chiar îmbunătăţi. Nu ştim încă dacă noi, oamenii, suntem singurii privitori de stele din cosmos, nici măcar dacă suntem primii, dar am aflat deja suficiente lucruri despre Univers pentru a şti că are potenţialul de a se trezi mult mai deplin decât s-a trezit până acum. Pionieri din domeniul inteligenţei artificiale precum Norbert Wiener ne-au învăţat că o trezire suplimentară a capacităţii Universului nostru de a procesa şi experimenta informaţii nu necesită neapărat eoni de evoluţie suplimentară, ci poate că doar câteva decenii de ingeniozitate ştiinţifică omenească.

Am putea fi ca acea primă sclipire de conştiinţă de sine pe care ai trăit-o când te-ai trezit din somn azi de dimineaţă, o premoniţie a conştiinţei mult superioare că va urma când deschizi ochii şi te trezeşti de-a binelea. Poate că superinteligenţa artificială va permite vieţii să se infiltreze în tot cosmosul şi să prospere miliarde sau trilioane de ani, şi poate că acest lucru se va întâmpla datorită unor decizii pe care le vom lua aici, pe planeta noastră, în decursul vieţii noastre.

Sau poate că umanitatea va pieri printr-o calamitate determinată de ea însăşi, cauzată de puterea tehnologiei care creşte mai repede decât înţelepciunea cu care o administrăm.

Dezbaterea în desfăşurare cu privire la impactul inteligenţei artificiale

asupra societăţii

Mulţi gânditori resping ideea de superinteligenţă ca fiind science-fiction, deoarece văd inteligenţa ca pe ceva misterios ce poate exista doar în organisme biologice – în special în oameni – şi ca fiind fundamental limitată la ceea ce pot face oamenii de astăzi. Dar, din perspectiva mea de fizician, inteligenţa nu este decât un tip de prelucrare a informaţiei realizat de particule elementare care se mişcă, şi nu există nicio lege a fizicii care să spună că cineva nu poate construi maşini mai inteligente decât noi din orice punct de vedere, putând astfel să însămânţeze viaţa cosmică. Acest lucru sugerează că am văzut doar vârful aisbergului inteligenţei; există un potenţial uimitor pentru deblocarea întregii inteligenţe latente din natură şi de a o utiliza pentru a ajuta umanitatea să înflorească sau să se ofilească.

Alţii, printre care se numără şi câţiva dintre autorii acestui volum, nu cred în construirea unei inteligenţe artificiale generale (entitate ce poate efectua orice sarcină cognitivă cel puţin la fel de bine ca oamenii), nu pentru că o consideră fizic imposibilă, ci pentru că o consideră prea dificil de obţinut pentru oameni în mai puţin de un secol. Printre cercetătorii profesionişti din domeniul inteligenţei artificiale, ambele tipuri de respingere au devenit perspective minoritare datorită descoperirilor recente. Mulţi oameni se aşteaptă ca inteligenţa artificială generală să fie obţinută într-un secol, iar prognoza medie este în doar câteva decenii. Un studiu recent al cercetătorilor din domeniul inteligenţei artificiale realizat de Vincent Müller şi Nick Bostrom conchide:

> *Rezultatele arată părerea experţilor conform căreia sistemele de inteligenţă artificială probabil (peste 50%) vor egala capacităţile generale ale omului până în anii '2040 – 2050 şi foarte probabil (cu o probabilitate de 90%) până în anul 2075. De la atingerea capacităţilor omeneşti, se va ajunge la superinteligenţă în 2 ani (10%) până la 30 de ani (75%) ulterior.*[29]

În perspectiva cosmică de giga-ani, nu contează prea mult dacă inteligenţa artificială generală soseşte în 30 sau în 300 de ani, aşa că haideţi să ne concentrăm pe consecinţe, şi nu pe timp.

În primul rând, noi, oamenii, am descoperit cum să replicăm anumite procese naturale cu maşinile, făcându-ne propria căldură, lumină şi putere mecanică. Gradual, am observat că şi corpurile noastre sunt maşini şi descoperirea celulelor nervoase a înceţoşatgraniţele dintre corp şi minte. În cele din urmă, am început să construim maşini care să ne întreacă atât muşchii, cât şi minţile, în prezent, suntem eclipsaţi de maşini în realizarea multor sarcini cognitive restrânse, de la memorare şi aritmetică până la jocuri, şi suntem pe cale să fim întrecuţi în multe alte privinţe, de la condus

până la investiţii şi diagnosticare medicală. În cazul în care comunitatea inteligenţei artificiale reuşeşte să-şi atingă obiectivul său iniţial de a construi o inteligenţă artificială generală, atunci vom fi, prin definiţie, eclipsaţi în ceea ce priveşte toate sarcinile noastre cognitive.

Acest lucru naşte multe întrebări evidente. De exemplu, cine va controla inteligenţa artificială generală va controla şi Pământul? Ar trebui să încercăm să controlăm maşinile superinteligente? Dacă nu, ne putem asigura că înţeleg, adoptă şi păstrează valorile umane? Aşa cum spunea Norbert Wiener în *The Human Use of Human Beings*:

> *„Vai de noi dacă lăsăm [maşina] să ne decidă comportamentul înainte de a examina legile acţiunilor sale şi de a şti bine dacă acest comportament al maşinii va fi ghidat de principii acceptabile pentru noi! Pe de altă parte, maşina […] care poate învăţa şi lua decizii în baza a ceea ce a învăţat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi".*

Şi cine suntem „noi" în această ecuaţie? Cine stabileşte dacă „aceste decizii sunt acceptabile"? Chiar dacă puterile viitoare decid să-i ajute pe oameni să supravieţuiască şi să prospere, cum vom găsi sens şi scop în vieţile noastre dacă nu mai este nevoie de noi pentru nimic?

Dezbaterea cu privire la impactul inteligenţei artificiale asupra societăţii s-a schimbat semnificativ în ultimii câţiva ani. În 2014, puţinele discuţii despre riscurile inteligenţei artificiale aveau tendinţa să fie catalogate drept încercări lucidite de creare a fricii, pentru unul dintre următoarele două motive incompatibile:

1. *Inteligenţa artificială generală era prezentată exagerat şi nu urma să apară timp de cel puţin încă un secol.*
2. *Inteligenţa artificială generală putea să apară mai devreme, dar aproape garantat va fi benefică.*

Astăzi discuţiile despre impactul inteligenţei artificiale asupra societăţii sunt peste tot şi activităţile ce vizează siguranţa şi etica inteligenţei artificiale au fost mutate în companii, universităţi şi conferinţe academice. Poziţia controversată a cercetării siguranţei inteligenţei artificiale nu mai înseamnă susţinerea, ci negarea ei. Întrucât scrisoarea deschisă care a reieşit dintr-o conferinţă despre inteligenţa artificială ce a avut loc în Porto Rico în anul 2015 (şi a ajutat la sporirea siguranţei inteligenţei artificiale) vorbea doar în termeni vagi despre importanţa de a păstra benefică inteligenţa artificială, principiile Asilomar privind inteligenţa artificială (vezi mai jos) aveau dinţi ascuţiţi: menţionează explicit autoîmbunătăţirea,

superinteligenţa şi riscul existenţial, fiind semnate de lideri din domeniul inteligenţei artificiale şi de peste o mie de cercetători din domeniu din întreaga lume.

Cu toate acestea, cele mai multe discuţii sunt limitate la impactul pe termen scurt al inteligenţei artificiale concentrate, iar comunitatea mai largă acordă doar o atenţie limitată transformărilor dramatice pe care inteligenţa artificială generală le poate aduce în curând vieţii de pe Pământ. De ce?

De ce ne grăbim să devenim depăşiţi şi de ce evităm să vorbim despre asta

În primul rând, din motive simple de economie. De câte ori găsim o cale nouă de a face munca omenească depăşită, construind maşini care o fac mai bine şi mai ieftin, cea mai mare parte a societăţii câştigă: cei care construiesc şi utilizează maşinile fac profit, iar consumatorii obţin produse mai ieftine. Acest lucru va fi la fel de adevărat şi pentru inteligenţele artificiale generale cu rol de investitor sau inteligenţele artificiale generale cu rol de cercetător, cum a fost şi pentru maşinile de ţesut, excavatoare şi roboţii industriali. În trecut, angajaţii scoşi de pe piaţa muncii şi-au găsit alte meserii, dar acest stimulent economic elementar va rămâne chiar dacă nu mai este cazul. Existenţa inteligenţelor artificiale generale ieftine înseamnă, prin definiţie, că *toate* meseriile vor putea fi făcute mai ieftin de maşini, deci oricine pretinde că „oamenii vor găsi întotdeauna meserii noi şi bine plătite" pretinde, de fapt, că cercetătorii din domeniul inteligenţei artificiale nu vor reuşi să construiască o inteligenţă artificială generală.

În al doilea rând, *Homo sapiens* este curios de la natură, calitate ce va motiva obiectivul ştiinţific de a înţelege inteligenţa şi de a dezvolta o inteligenţă artificială generală chiar şi fără stimulente economice. Deşi curiozitatea este una dintre cele mai slăvite calităţi omeneşti, poate cauza probleme când implică tehnologii pe care nu am învăţat cum să le stăpânim înţelept. Curiozitatea ştiinţifică pură fără motivaţia profitului a contribuit la descoperirea armelor nucleare şi instrumentelor de inginerie a virusurilor, deci nu este de neconceput ca vechea zicală „curiozitatea a omorât pisica" să se aplice şi speciei umane, în cele din urmă.

În al treilea rând, suntem muritori. Acest lucru explică suportul aproape unanim pentru dezvoltarea noilor tehnologii care ne ajută să trăim vieţi mai lungi şi mai sănătoase, ceea ce motivează puternic cercetările prezente din domeniul inteligenţei artificiale. Inteligenţa artificială generală poate ajuta cercetarea medicală chiar şi mai mult. Anumiţi gânditori chiar speră să obţină cvasinemurirea prin cyborgizare sau încărcare.

Astfel ne aflăm pe o pantă alunecoasă către inteligenţa artificială generală, cu stimulente puternice pentru a continua să alunecăm în jos, deşi consecinţa va fi, prin definiţie, depăşirea noastră economică. Nu va mai fi nevoie de noi pentru nimic, pentru că toate meseriile vor putea fi făcute mai bine de maşini. Crearea reuşită a inteligenţei artificiale generale ar fi cel mai mare eveniment din istoria umanităţii, aşadar de ce există atât de puţine discuţii serioase cu privire la unde ne va conduce acest lucru?

Şi aici răspunsul implică motive multiple.

În primul rând, cum remarca ironic Upton Sinclair, „este dificil să-l faci pe om să înţeleagă ceva când salariul său depinde de a nu înţelege”[30]. De exemplu, purtătorii de cuvânt ai companiilor de tehnologie sau ai grupurilor de cercetare universitară pretind adesea că nu există riscuri în activităţile lor chiar dacă în particular gândesc altfel. Observaţia lui Sinclair poate ajuta la explicarea nu doar a reacţiilor la riscuri determinate de fumat şi încălzirea globală, ci şi la motivele pentru care unii tratează tehnologia ca pe o nouă religie ale cărei credinţe centrale sunt că mai multă tehnologie este întotdeauna mai bine şi ai cărei eretici sunt luddiţi temători şi care nu o înţeleg.

În al doilea rând, oamenii au o istorie veche cu autoiluzionarea, extrapolarea incorectă a trecutului şi subestimarea tehnologiilor emergente. Evoluţia darwinistă ne-a dotat cu o frică puternică de ameninţări concrete, nu de ameninţările abstracte ale tehnologiilor viitoare pe care ne este greu să le vizualizăm sau chiar să ni le imaginăm. Gândeşte-te cum ar fi fost să încerci să-i avertizezi pe oamenii din anii 1930 de o cursă a înarmărilor nucleare viitoare, când nu le-ai fi putut arăta un singur clip video cu o explozie nucleară şi nimeni nu ştia măcar cum se fabrică astfel de arme. Chiar şi cercetătorii de top pot subestima nesiguranţa, făcând prognoze care sunt fie prea optimiste – unde sunt reactoarele de fuziune şi maşinile zburătoare? –, fie prea pesimiste. Ernest Rutherford, probabil cel mai mare fizician nuclear al vremii sale, spunea în 1933 – cu mai puţin de douăzeci şi patru de ore înainte ca Leo Szilard să conceapă reacţia nucleară în lanţ – că energia nucleară este „o prostie”. Efectiv, nimeni din vremea aceea nu a putut anticipa cursa înarmărilor nucleare.

În al treilea rând, psihologii au descoperit că avem tendinţa să evităm să ne gândim la ameninţări neliniştitoare când credem că nu putem face nimic pentru a le evita. Însă, în acest caz, sunt multe lucruri constructive pe care le putem face, dacă am putea să începem să ne gândim serios la problemă.

Ce putem face?

Susţin schimbarea strategiei de la „hai să ne grăbim să construim tehnologii care să ne facă să devenim depăşiţi – ce se poate

întâmpla?” la „să ne imaginăm un viitor luminos şi să facem eforturi să-l transformăm în realitate”.

Pentru a motiva efortul necesar acestei transformări, strategia începe prin imaginarea unei destinaţii ispititoare. Deşi viitorurile portretizate de Hollywood au tendinţa să fie distopice, adevărul este că inteligenţa artificială generală poate ajuta viaţa să înflorească mai mult decât oricând. Tot ce iubesc eu la civilizaţie este produsul inteligenţei, deci, dacă putem să ne amplificăm inteligenţa prin inteligenţă artificială generală, avem posibilitatea de a rezolva cele mai spinoase probleme de astăzi şi de mâine, inclusiv bolile, încălzirea globală şi sărăcia. Cu cât vom face mai detaliată viziunea noastră comună şi pozitivă despre viitor, cu atât vom fi mai motivaţi să lucrăm împreună pentru a o transforma în realitate.

Ce ar trebui să facem pentru a ghida această viziune? Cele douăzeci şi trei de principii Asilomar adoptate în 2017 oferă suficientă ghidare, inclusiv următoarele obiective pe termen scurt:

1. *O cursă a înarmărilor cu arme autonome letale trebuie evitată.*
2. *Prosperitatea economică creată de inteligenţa artificială trebuie distribuită la scară largă, astfel încât întreaga umanitate să se bucure de ea.*
3. *Investiţiile în inteligenţa artificială trebuie însoţite de fonduri de cercetare speciale destinate asigurării utilizării ei benefice… Cum putem face sistemele viitoare de inteligenţă artificială foarte robuste, astfel încât să poată face ce vrem noi fără să dea erori de funcţionare sau să fie atacate de hackeri?*[31]

Primele două implică să nu rămânem blocaţi în echilibre Nash suboptime. O cursă a înarmărilor cu arme autonome letale scăpată de sub control, care determină scăderea continuă a preţurilor asasinatelor anonime automate până la zero, va fi foarte greu de oprit dacă prinde inerţie. Al doilea scop ar necesita inversarea tendinţei moderne în câteva ţări occidentale în care sectoare întregi de populaţie sunt din ce în ce mai sărace în termeni de sărăcie absolută, alimentând astfel furia, resentimentele şi polarizarea. Dacă al treilea obiectiv nu este realizat, toată tehnologia minunată a inteligenţei artificiale pe care o creăm ne-ar putea vătăma, fie accidental, fie intenţionat.

Cercetarea siguranţei inteligenţei artificiale trebuie făcută cu un termen-limită strict în minte: înainte de sosirea inteligenţei artificiale, trebuie să stabilim cum o vom face să înţeleagă, adopte şi păstreze obiectivele noastre. Cu cât maşinile devin mai puternice şi mai inteligente, cu atât devine mai importantă alinierea obiectivelor lor cu ale noastre. Cât timp construim maşini relativ lipsite de inteligenţă, întrebarea nu este dacă obiectivele omeneşti vor prevala, ci câte

probleme ne-ar putea cauza maşinile înainte de a putea rezolva problema alinierii obiectivelor. Însă, dacă o superinteligenţă este vreodată eliberată, va fi invers: deoarece inteligenţa este capacitatea de a realiza obiective, o inteligenţă artificială superinteligenţă este prin definiţie mult mai bună în realizarea obiectivelor sale decât sunt oamenii la realizarea obiectivelor lor şi, prin urmare, va triumfa.

În alte cuvinte, *riscul real al inteligenţei artificiale generale nu este rea-voinţa ci competenţa*. O inteligenţă artificială generală superinteligenţă va fi extrem de bună la realizarea obiectivelor sale şi, dacă acele obiective nu sunt aliniate cu ale noastre, am încurcat-o. Oamenii nu se gândesc prea mult când inundă muşuroaiele pentru a construi diguri hidroelectrice, aşa că am face bine să nu punem umanitatea în locul acelor furnici. Majoritatea cercetătorilor argumentează că, dacă vom crea până la urmă o superinteligenţă, ar trebui să ne asigurăm că este ceea ce pionierul în domeniul siguranţei inteligenţei artificiale Eliezer Yudkowsky a numit „inteligenţă artificială prietenoasă" – o inteligenţă artificială ale cărei obiective sunt benefice într-un sens profund.

Problema morală legată de ce ar trebui să fie aceste obiective este la fel de urgentă ca problemele tehnice legate de alinierea obiectivelor. De exemplu, ce fel de societate sperăm să creăm, unde găsim sens şi scop în vieţile noastre chiar dacă noi, la drept vorbind, nu mai suntem necesari? Adesea primesc următorul răspuns uşuratic la această întrebare: „Să construim maşini mai inteligente decât noi şi să le lăsăm pe ele să găsească răspunsul!" Acest răspuns pune greşit semnul de egalitate între inteligenţă şi moralitate. Inteligenţa nu este bună sau rea, ci neutră din punct de vedere moral. Este pur şi simplu capacitatea de a realiza obiective complexe, bune sau rele. Nu putem trage concluzia că lucrurile ar fi stat mai bine dacă Hitler ar fi fost mai inteligent. Într-adevăr, amânarea abordării problemelor de etică până după construirea inteligenţei artificiale generale cu obiective aliniate este iresponsabilă şi potenţial dezastruoasă. O superinteligenţă perfect obedientă ale cărei obiective se aliniază automat cu cele ale proprietarilor umani ar fi un fel de SS-Obersturmbannführer Adolf Eichmann nazist pe steroizi. În lipsa unei busole morale sau a unor inhibiţii personale, ar implementa cu o eficienţă nemiloasă obiectivele proprietarilor săi, oricare ar fi acestea.[32]

Când vorbesc despre nevoia de a analiza riscurile tehnologiei, sunt acuzat că încerc să stârnesc teama. Dar aici, la MIT, unde lucrez, ştim că o astfel de analiză a riscurilor nu este o încercare de a stârni teama: este *ingineria siguranţei*. Înainte de misiunea de aterizare pe Lună, NASA s-a gândit sistematic la tot ce ar putea să nu meargă când vor pune astronauţi în vârful unei rachete de 110 metri, plină cu combustibil foarte inflamabil, şi-i va lansa către un loc unde nimeni

nu-i putea ajuta – şi erau multe lucruri care puteau să nu meargă. Era stârnire a fricii? Nu, era ingineria siguranţei care a asigurat succesul misiunii. Similar, ar trebui să analizăm ce ar putea să nu meargă bine cu inteligenţa artificială pentru a ne asigura că va merge bine.

Perspectivă

Pe scurt, dacă tehnologia depăşeşte înţelepciunea cu care o putem gestiona, ne poate conduce la extincţie. Deja a determinat extincţia a 20 până la 50 la sută din totalitatea speciilor de pe Pământ, după unele estimări[33], şi ar fi o ironie a sorţii dacă noi am fi următorii. În acelaşi timp, ar fi jalnic, dat fiind faptul că oportunităţile oferite de inteligenţa artificială generală sunt practic astronomice, oferind potenţial vieţii să prospere miliarde de ani nu doar pe Pământ, ci şi într-o mare parte din cosmosul nostru.

În loc să irosim această oportunitate printr-o negare neştiinţifică a riscurilor şi printr-o planificare proastă, haideţi să fim ambiţioşi! *Homo sapiens* este exaltant de ambiţios, cum spun şi faimoasele versuri ale lui William Ernest Henley din „Invictus": „Eu, azi, al sorţii mele sunt stăpânul, / Sunt căpitanul sufletului meu"[34]. În loc să ne lăsăm duşi de val ca o corabie fără cârmă către propria noastră depăşire, să abordăm şi să depăşim provocările tehnice şi sociale care stau între noi şi un viitor tehnologizat bun. Dar cum rămâne cu provocările existenţiale legate de moralitate, obiective şi sens? Nu există niciun sens codat în legile fizicii, aşa că, în loc să aşteptăm pasiv ca Universul să ne confere nouă sens, să acceptăm şi să sărbătorim faptul că noi, fiinţele conştiente, suntem cele care dăm sens Universului. Să ne creăm propriul sens, bazat pe ceva mai profund decât a avea pur şi simplu o meserie. Inteligenţa artificială generală ne poate permite să devenim, în sfârşit, stăpânii propriului nostru destin. Să facem acest destin unul cu adevărat înălţător!

Capitolul 9
Mesaje disidente

Jaan Tallinn

Jaan Tallinn, programator, fizician teoretician şi investitor, este cofondator Skype şi Kazaa.

Jaan Tallinn a crescut în Estonia, devenind unul dintre puţinii dezvoltatori de jocuri pe calculator de acolo, când acea naţiune era încă o republică socialistă sovietică. Aici, compară disidenţii care au făcut să cadă Cortina de Fier cu disidenţii care dau semnalul de alarmă pentru progresele rapide ale inteligenţei artificiale.

Plasează rădăcinile disidenţei curente din domeniul inteligenţei artificiale, paradoxal, în rândul unor pionieri ai domeniului inteligenţei artificiale precum Wiener, Alan Turing şi I.J. Good.

Preocuparea lui Jaan este cu riscurile existenţiale, inteligenţa artificială fiind printre cele mai extreme din mai multe. În 2012, a coanfiinţat Centrul pentru Studierea Riscurilor Existenţiale – un institut de cercetare interdisciplinară care lucrează la ameliorarea riscurilor „asociate tehnologiilor emergente şi activităţii umane" – la Universitatea Cambridge, împreună cu filosoful Huw Price şi Martin Rees, astronom regal.

S-a descris ca fiind „un consecinţialist convins" – suficient de convins încât să-şi dea o mare parte din bogăţia sa antreprenorială Institutului Future of Life (al cărui cofondator este), Machine Intelligence Research Institute şi altor astfel de organizaţii care lucrează la reducerea riscurilor. Maxâegmark a scris despre el: „Dacă eşti o formă de viaţă inteligentă care citeşte acest text peste câteva milioane de ani şi eşti uimit de cum prosperă viaţa, probabil îţi datorezi existenţa lui Jaan".

Într-o vizită recentă la Londra, eu şi Jaan am participat la o conferinţă pe tema inteligenţei artificiale, pentru maratonul Serpentine Gallery organizat de Primăria Londrei, sub egida lui Hans Ulrich Obrist (alt participant la acest volum). Fiind vorba despre lumea artistică, a fost o cină opulentă într-o vilă plină cu oamenii frumoşi ai Londrei: artişti, fotomodele, oligarhi, vedete de teatru şi de cinema. După ce s-a prezentat în felul său nepretenţios („Salut, sunt Jaan"), a spus dintr-odată, „este timpul să dansăm pe hip-hop", s-a aruncat pe podea într-o mână şi a început să-şi etaleze mişcările spectaculoase în faţa personalităţilor năucite din încăpere. Apoi a mers în subcultura cluburilor de dans, aparent acesta fiind modul în care îşi încheie fiecare seară când este în turneu. Cine ar fi ştiut?

În martie 2009, eram într-un restaurant fad de pe marginea unei autostrăzi gălăgioase din California. Eram acolo pentru a mă întâlni cu un tânăr al cărui blog îl urmăream. Pentru a se face recunoscut, purta un nasture pe care scria: *spune adevărul chiar dacă îţi tremură vocea*. Numele său era Eliezer Yudkowsky şi ne-am petrecut următoarele patru ore discutând despre mesajul pe care îl are el pentru lume – mesajul care mă adusese în acel restaurant şi urma să domine activitatea mea ulterioară.

Primul mesaj: ocupaţia sovietică

În *The Human Use of Human Beings*, Norbert Wiener a privit lumea prin lentila comunicaţiei. A văzut un Univers care mergea în paşi de

dans pe ritmul celei de-a doua legi a termodinamicii către inevitabila sa moarte termică. Într-un astfel de Univers, singurele entităţi (meta) stabile sunt *mesajele* – tipare de informaţie ce se propagă prin timp, ca undele care se propagă pe suprafaţa unui lac. Chiar şi noi, oamenii, putem fi consideraţi mesaje, deoarece atomii din corpurile noastre sunt prea trecători pentru a ne ataşa identităţile noastre lor. În schimb, suntem „mesajul" pe care îl întreţin funcţiile noastre corporale. După cum spune Wiener: „tiparul menţinut de această homeostază este piatra de hotar a identităţii noastre personale".

Sunt obişnuit să tratez procesele şi calculele ca pe nişte elemente fundamentale ale lumii. Acestea fiind spuse, lentila lui

Wiener scoate în evidenţă câteva aspecte interesante ale lumii care altfel ar fi rămas probabil în fundal şi care mi-au conturat viaţa, într-o mare măsură. Acestea sunt două mesaje, ambele avându-şi rădăcinile în cel de-al Doilea Război Mondial. Au început ca fiind mesaje disidente liniştite – mesaje cărora omenii nu le acordau prea mare atenţie, chiar dacă erau de acord cu ele tacit sau poate subconştient. Primul mesaj era: *Uniunea Sovietică este compusă dintr-o serie de ocupaţii ilegitime. Aceste ocupaţii trebuie să înceteze*.

Eston fiind, am crescut în spatele Cortinei de Fier şi am fost în rândul întâi când a căzut. Am auzit acest prim mesaj în reminiscenţele nostalgice ale bunicilor mei şi între zgomotele aspre care bruiau Vocea Americii. Au fost şi mai puternice în timpul erei Gorbaciov, când statul a devenit mai blând în tratamentul aplicat disidenţilor, crescând apoi în timpul Revoluţiei Cântate din Estonia, la sfârşitul anilor 1980.

În adolescenţa mea, am fost martor la răspândirea acestui mesaj în cercuri din ce în ce mai mari de oameni, începând cu disidenţii activi, care îl vociferau de jumătate de secol şi acest lucru îi costa mult, continuând cu artiştii şi oamenii de litere şi sfârşind în rândurile membrilor de partid şi politicienilor care au schimbat tabăra. Această nouă elită consta într-un grup eclectic de oameni: disidenţii iniţiali care au reuşit să supravieţuiască represiunii, intelectualii publici şi (spre exasperarea disidenţilor supravieţuitori) chiar foşti comunişti. Dogmaticii rămaşi – chiar şi cei remarcabili – au fost marginalizaţi, în cele din urmă, câţiva dintre ei retrăgându-se în Rusia.

Interesant, pe măsură ce mesajul se propaga de la un grup la altul, evolua. A început într-o formă pură şi intransigentă („ocupaţia trebuie să ia sfârşit!") în rândurile disidenţilor care considerau că adevărul este mai important decât libertatea lor personală. Grupurile populare, care aveau mai mult de pierdut, au calificat şi au diluat iniţial mesajul, luând poziţii ca „ar fi bine pe termen lung să delegăm controlul asupra problemelor locale" (Apăreau întotdeauna excepţii: unii intelectuali publici proclamau mesajul disident iniţial *ad litteram*). În cele din urmă, mesajul original – fiind pur şi simplu adevărat – a

prevalat asupra versiunilor sale diluate. Estonia şi-a recâştigat independenţa în 1991 şi ultimele trupe sovietice au plecat după trei ani.

Oamenii care şi-au asumat riscuri şi au grăit adevărul în Estonia şi în alte ţări din blocul sovietic au jucat un rol monumental în rezultatul final – un rezultat care a schimbat vieţile a sute de milioane de oameni, inclusiv a mea. Au grăit adevărul, chiar dacă le tremurau vocile.

Al doilea mesaj: riscul inteligenţei artificiale

Expunerea mea la cel de-al doilea mesaj revoluţionar a fost prin intermediul blogului lui Yudkowsky – blogul care m-a făcut să-l contactez şi să mă întâlnesc cu el în California. Mesajul era: *Progresul continuu în domeniul inteligenţei artificiale poate precipita o schimbare de proporţii cosmice – un proces scăpat de sub control care prezintă posibilitatea de a omorî pe toată lumea. Trebuie să facem foarte mult efort suplimentar pentru a evita acest rezultat.*

După întâlnirea mea cu Yudkqwsky, primul lucru pe care l-am făcut a fost să încerc să stârnesc interesul colegilor mei de la Skype şi colaboratorilor apropiaţi cu privire la această avertizare. Nu am reuşit. Mesajul era prea nebunesc, prea disident. Vremea sa nu venise încă.

Abia mai târziu am aflat că Yudkowsky nu era primul disident care a grăit acest adevăr. În aprilie 2000, în revista *Wired* a apărut un articol de opinie lung intitulat „De ce viitorul nu are nevoie de noi", scris de Bill Joy, cofondatorul şi directorul ştiinţific al companiei Sun Microsystems. El avertiza:

> *„Obişnuiţi a trăi cu descoperirile ştiinţifice revoluţionare, nu ne-am împăcat încă cu faptul că cele mai interesante tehnologii ale secolului al XXI-lea – robotica, ingineria genetică şi nanotehnologia – reprezintă o ameninţare diferită faţă de tehnologiile anterioare. În special roboţii, organismele create pe cale inginerească şi nanoboţii au în comun un factor amplificator periculos: se pot replica automat… un «bot» se poate înmulţi şi lucrurile pot scăpa rapid de sub control".*

Se pare că atacul lui Joy a făcut furori, dar nu a stârnit nicio acţiune.

Însă pentru mine a fost mai surprinzător că mesajul privind riscurile inteligenţei artificiale a luat naştere aproape simultan cu domeniul informaticii. Într-o prelegere din 1951, Alan Turing anunţa: „Pare probabil ca, odată ce am iniţiat gândirea automată, să nu mai dureze mult până ne va depăşi puterile noastre plăpânde… Aşadar, la un anumit nivel, ar trebui să ne aşteptăm ca maşinile să preia controlul…"[35] După aproximativ un deceniu, colegul său din Bletchley

Park, I.J. Good, scria: „prima maşină ultrainteligentă este *ultima* invenţie pe care trebuie să o facă omul, dacă maşina este suficient de docilă încât să ne spună cum o putem ţine sub control”[36]. Într-adevăr, am numărat şase locuri în *The Human Use of Human Beings* în care Wiener a făcut aluzie la un aspect sau altul al Problemei Controlului. („Maşina asemănătoare djinnului, care poate învăţa şi poate lua decizii în baza învăţăturii sale, nu va fi obligată în niciun fel să ia deciziile pe care ar fi trebuit să le luăm noi sau care ar fi acceptabile pentru noi.”) Aparent, disidenţii iniţiali care promulgau mesajul despre riscurile inteligenţei artificiale sunt chiar pionierii inteligenţei artificiale!

Greşeala fatală a evoluţiei

Au existat multe argumente, unele sofisticate, altele mai puţin, pentru faptul că Problema Controlului este reală, şi nu o fantezie science-fiction. Permiteţi-mi să vă prezint una care ilustrează magnitudinea problemei:

În ultima sută de mii de ani, lumea (însemnând Pământul, dar argumentul se poate extinde şi la sistemul solar şi, probabil, chiar la întregul Univers) a fost în regimul creierului omenesc. În acest regim, creierele fiinţelor *Homo sapiens* au fost cele mai sofisticate mecanisme de conturare a viitorului (unii le-au numit cele mai complicate obiecte din univers). Iniţial, nu le-am folosit la multe altele decât pentru supravieţuire şi politică tribală într-un grup de culegători, însă acum efectele lor le depăşesc pe cele ale evoluţiei naturale. Planeta a trecut de la producţia de păduri la producţia de oraşe.

Aşa cum a prezis Turing, când vom avea o inteligenţă artificială supraomenească („metoda gândirii automate”), regimul creierului omenesc se va sfârşi. Priveşte în jurul tău – eşti martorul ultimelor decenii dintr-un regim de o sută de mii de ani. Numai acest gând ar trebui să-i facă pe oameni să mediteze înainte de a expedia inteligenţa artificială doar ca pe un alt instrument. Unul dintre cei mai importanţi cercetători din domeniul inteligenţei artificiale mi-a mărturisit că s-ar simţi foarte uşurat să audă că inteligenţa artificială de nivel omenesc este imposibil de creat.

Bineînţeles, probabil va mai dura mult timp până vom dezvolta inteligenţă artificială la nivel omenesc. Dar avem motive să bănuim că nu aşa stau lucrurile. La urma urmei, nu a durat mult, în termeni relativi, ca evoluţia – procesul de optimizare orb şi stângaci – să creeze inteligenţa omenească după ce a avut animalele ca materie primă. Sau chiar viaţa pluricelulară: se pare că i-a fost mult mai greu să facă celulele să se alipească decât i-a fost să creeze inteligenţa omenească după ce a avut organismele pluricelulare. Fără să mai menţionăm faptul că nivelul nostru de inteligenţă era limitat de factori

groteşti precum lăţimea canalului de naştere. Imaginează-ţi un dezvoltator al inteligenţei artificiale care nu mai poate avansa pentru că nu poate ajusta mărimea fontului pe calculatorul său!

Există aici o simetrie interesantă: în producţia de oameni, evoluţia a creat un sistem care este, cel puţin în multe dimensiuni importante, un planificator şi un organizator mai puternic decât evoluţia însăşi. Suntem prima specie care a înţeles că este produsul evoluţiei. În plus, am creat multe artefacte (radiouri, arme de foc, nave spaţiale) pe care evoluţia avea slabe şanse să le creeze. Prin urmare, viitorul nostru va fi determinat de propriile noastre decizii, şi nu de evoluţia noastră biologică. În acest sens, evoluţia a căzut victimă propriei sale Probleme a Controlului.

Putem doar să sperăm că suntem mai inteligenţi decât evoluţia în această privinţă. *Suntem* mai inteligenţi, bineînţeles, dar va fi oare suficient? Suntem pe cale de a afla.

Situaţia curentă

Aşadar, iată-ne, la peste jumătate de secol după avertizările iniţiale ale lui Turing, Wiener şi Good, şi la un deceniu după ce oameni ca mine au început să acorde atenţie mesajului privind riscurile inteligenţei artificiale. Mă bucur să văd că am făcut multe progrese în confruntarea acestor probleme, dar cu siguranţă nu le-am dat de cap încă. Riscurile inteligenţei artificiale, deşi nu mai sunt un subiect tabu, nu sunt încă apreciate la valoarea lor reală în rândurile cercetătorilor din domeniu. Riscurile inteligenţei artificiale nu sunt nici lucruri de care să ştie toată lumea. În raport cu cronologia primului nostru mesaj disident, aş spune că suntem aproximativ în anul 1988, când pornirea unei discuţii despre ocupaţia sovietică nu mai era ceva ce ţi-ar putea distruge cariera, dar trebuia să-ţi acoperi totuşi cumva poziţia. Aud acoperiri similare şi în ziua de azi – afirmaţii precum „nu-mi fac probleme legate de o superinteligenţă artificială, dar există probleme etice reale determinate de gradul din ce în ce mai mare de automatizare” sau „este bine că unii oameni cercetează şi riscurile inteligenţei artificiale, dar nu este o îngrijorare pe termen scurt” sau chiar ceva ce sună rezonabil precum „acestea sunt scenarii cu o probabilitate foarte mică, dar potenţialul impact major justifică atenţia acordată”.

Cât despre propagarea mesajului, ne apropiem de punctul de cotitură. Un studiu recent realizat pe cercetătorii inteligenţei artificiale care au publicat la cele două conferinţe internaţionale majore pe tema inteligenţei artificiale din 2015 arată că 40% dintre aceştia consideră acum că riscurile unei inteligenţe artificiale superavansate sunt fie „o problemă importantă”, fie „una dintre cele mai importante probleme din domeniu”[37]

Bineînţeles, la fel cum au existat comunişti dogmatici care nu şi-au schimbat niciodată poziţia, este garantat şi faptul că anumiţi oameni nu vor accepta niciodată ideea că inteligenţa artificială poate fi periculoasă. Mulţi dintre cei care au negat comunismul veneau din nomenclatura sovietică; similar, cei care neagă riscurile inteligenţei artificiale au adesea motive financiare sau alte motive pragmatice. Unul dintre principalele motive este reprezentat de profiturile corporaţiilor. Inteligenţa artificială este profitabilă şi, chiar şi în situaţiile în care nu este, este cel puţin o întreprindere trendy, orientată către viitor, cu care să-ţi asociezi compania. Astfel multe dintre poziţiile de negare sunt produse ale departamentului de relaţii publice din companii şi maşinaţii juridice. Într-un sens foarte real, marile corporaţii sunt maşini nonumane care-şi urmăresc propriile interese – interese care pot să nu fie aliniate cu cele ale oamenilor care lucrează pentru ele. Aşa cum observa Wiener în *The Human Use of Human Beingsr* „Când atomii omeneşti sunt croşetaţi într-o organizaţie în care sunt utilizaţi, nu în dreptul lor deplin de fiinţe umane responsabile, ci ca rotiţe, tije şi manete, contează prea puţin că materia lor primă este din came şi sânge”.

Un alt stimulent puternic pentru a închide ochii la riscurile inteligenţei artificiale este curiozitatea (foarte omenească) care nu cunoaşte nicio limită. „Când vezi ceva care este atrăgător din punct de vedere tehnic, îl transformi în realitate şi începi să discuţi despre ce ar trebui făcut doar după ce ai obţinut succesul tehnic. La fel au stat lucrurile şi cu bomba atomică”, a spus J. Robert Oppenheimer. Cuvintele sale au fost repetate recent de Geoffrey Hinton, despre care s-ar putea spune că este inventatorul învăţării profunde, în contextul riscurilor inteligenţei artificiale: „ţi-aş putea da argumentele obişnuite, dar adevărul este că ideea descoperirii este mult prea atrăgătoare”.

Incontestabil, putem mulţumi atitudinii antreprenoriale şi curiozităţii ştiinţifice pentru aproape toate lucrurile frumoase pe care le luăm de bune în epoca modernă. Însă, este important să ne dăm seama că progresul nu ne *datorează* un viitor bun. În cuvintele lui Wiener, „este posibil să crezi în progres ca stare de fapt fără să crezi în progres ca principiu etic”.

În cele din urmă, nu ne permitem luxul de a aştepta până când toţi corporatiştii şi cercetătorii inteligenţei artificiale vor fi de acord să accepte riscurile inteligenţei artificiale. Imaginează-ţi că stai într-un avion care urmează să decoleze. Dintr-odată, auzi un anunţ care spune că 40% dintre experţi consideră că la bord se află o bombă. În acel moment, e clar ce ai de făcut şi e clar că nu vei sta acolo să aştepţi să cadă de acord şi ceilalţi 60%.

Calibrarea mesajului privind riscurile

inteligenţei artificiale

Deşi este uimitor de profetic, mesajul privind riscurile inteligenţei artificiale venit de la disidenţii iniţiali are un mare defect – la fel ca versiunea care domină discursul public curent: ambele subestimează considerabil magnitudinea problemei, precum şi avantajele potenţiale ale inteligenţei artificiale. Mesajul, în alte cuvinte, nu transmite adecvat adevăratele mize ale jocului.

Wiener a avertizat iniţial în legătură cu riscurile *sociale* – riscurile care îşi vor avea originea în integrarea neglijentă a deciziilor generate de maşini cu procesele de guvernare şi proasta utilizare (de către oameni) a acestui proces automat de luare a deciziilor.

Asemenea, dezbaterea „serioasă" curentă cu privire la riscurile inteligenţei artificiale se concentrează în primul rând pe lucruri precum şomaj tehnologic sau prejudecăţi ale învăţării automate. Deşi astfel de discuţii pot fi valoroase şi pot aborda probleme serioase pe termen scurt, sunt extrem de înguste. Îmi aduc aminte de remarca lui Yudkowsky dintr-o postare pe blogul său: „să te întrebi care sunt efectele *superinteligenţei automate* asupra unei pieţe de muncă omeneşti convenţionale este ca şi cum te-ai întreba cum vor fi afectate tiparele de comerţ dintre SUA şi China dacă s-ar izbi Luna de Pământ. Cu siguranţă ar exista efecte, dar nu ai înţeles ideea principală".

Din punctul meu de vedere, punctul central al riscurilor inteligenţei artificiale este acela că *inteligenţa artificială superinteligentă este un risc de mediu*. Permiteţi-mi să explic.

În parabola sa despre „balta conştientă", Douglas Adams descrie o baltă ce se trezeşte într-o dimineaţă într-o groapă care i se potriveşte „surprinzător de bine". Din acea observaţie, balta trage concluzia că lumea trebuie să fie făcută pentru ea. Prin urmare, scrie Adams, „momentul în care dispare o ia pe nepregătite". Să presupunem că riscurile inteligenţei artificiale sunt limitate doar la dezvoltările sociale adverse înseamnă să facem o greşeală similară. Realitatea crudă este că universul nu este făcut pentru noi; în schimb, suntem ajustaţi de evoluţie pentru a supravieţui într-un interval foarte precis de parametri de mediu. De exemplu, avem nevoie ca atmosfera de la sol să fie aproximativ la temperatura camerei, să aibă aproximativ 100 kPa presiune şi o concentraţie suficientă de oxigen. Orice schimbare, chiar şi temporară, a acestui echilibru precar şi murim în câteva minute.

Inteligenţa bazată pe siliciu nu împărtăşeşte aceleaşi probleme legate de mediu. Acesta este motivul pentru care este mult mai ieftin să explorăm spaţiul cu sonde automate, şi nu cu „conserve de carne". În plus, mediul actual al Pământului este aproape sigur suboptim pentru ceea ce ar vrea o inteligenţă artificială superinteligentă: *putere de calcul*. Aşadar, ne putem trezi că planeta noastră trece de la o

încălzire globală antropogenetică la o răcire globală maşinogenetică. O provocare importantă pe care trebuie să o abordeze specialiştii în siguranţa inteligenţei artificiale este cum ar putea să ţină sub control o potenţială inteligenţă artificială superinteligentă – o inteligenţă artificială cu o amprentă mult mai mare decât a noastră –, astfel încât să nu transforme mediul nostru într-o zonă care nu mai poate fi locuită de forme de viaţă biologice.

În mod interesant, dat fiind faptul că cele mai puternice surse atât pentru cercetarea inteligenţei artificiale, cât şi pentru negarea riscurilor sale, sunt sub marile umbrele corporative, dacă-ţi încordezi privirea suficient, mesajul „inteligenţei artificiale ca risc de mediu" arată ca îngrijorarea cronică legată de corporaţiile care se eschivează în faţa responsabilităţii lor faţă de mediu.

În sensul invers, grijile despre efectele sociale ale inteligenţei artificiale nu percep nici părţile pozitive în totalitate. Este greu să exagerezi cât de mic şi de local este viitorul planetei noastre, comparat cu potenţialul deplin al umanităţii. La scări temporale astronomice, planeta noastră va dispărea în curând (dacă nu cumva învăţăm să îmblânzim Soarele, lucru care este şi el posibil) şi aproape toate resursele – atomi şi energie liberă – care vor susţine civilizaţia pe termen lung sunt în spaţiul adânc.

Eric Drexler, inventatorul nanotehnologiei, a popularizat recent conceptul de „pareto-topie": ideea conform căreia inteligenţa artificială, dacă este făcută bine, poate determina un viitor în care vieţile *tuturor* oamenilor sunt mult îmbunătăţite, un viitor în care nu mai există pierzători. O realizare-cheie în această privinţă este că principalul motiv pentru care umanitatea nu-şi atinge potenţialul deplin poate fi simţul nostru instinctiv că suntem într-un joc cu sumă zero – un joc în care jucătorii trebuie să obţină mici victorii unii pe seama altora. Un astfel de instinct este foarte dezorientat şi distructiv într-un „joc" unde miza este totul, iar câştigul potenţial este astronomic. Sunt mult mai multe sisteme solare în galaxia noastră decât sunt oameni pe Pământ.

Speranţa

Ca şi în această scriere, sunt un optimist prevăzător cu privire la faptul că mesajul riscurilor inteligenţei artificiale poate salva umanitatea de la extincţie, la fel cum mesajul ocupaţiei sovietice a sfârşit prin a elibera sute de milioane de oameni. Începând cu 2015, a atins şi a convertit 40% dintre cercetătorii inteligenţei artificiale. Nu m-ar surprinde dacă un nou studiu ar arăta că majoritatea cercetătorilor din domeniul inteligenţei artificiale consideră că siguranţa inteligenţei artificiale este o problemă importantă.

Sunt încântat să văd că urmează să apară primele lucrări tehnice

pe tema siguranţei inteligenţei artificiale de pe urma Deep Mind, OpenAI şi Google Brain şi a spiritului colaborator de soluţionare a problemelor, spirit care înfloreşte în echipele de cercetare a siguranţei inteligenţei artificiale din aceste organizaţii care altfel ar fi foarte competitive.

Şi elita politică şi comercială a lumii se trezeşte încet-încet: siguranţa inteligenţei artificiale a fost acoperită în rapoarte şi prezentări realizate de Institute of Electrical and Electronics Engineers (IEEE), World Economic Forum şi de Organization for Economic Cooperation and Development (OECD). Chiar şi recentul (iulie 2017) manifest chinez privind inteligenţa artificială conţine secţiuni speciale privind „supravegherea siguranţei inteligenţei artificiale", „dezvoltarea unor legi, reglementări şi norme etice" şi stabilirea unui „sistem de evaluare şi securitate a inteligenţei artificiale" pentru, printre altele, a „spori conştientizarea riscurilor". Sper din inimă ca noua generaţie de lideri care înţeleg problema controlului inteligenţei artificiale şi inteligenţa artificială ca risc ultim de mediu să depăşească obişnuitele jocuri tribale, cu sumă zero, şi să direcţioneze umanitatea dincolo de aceste ape periculoase în care ne aflăm – deschizându-ne astfel calea către stelele care ne-au aşteptat miliarde de ani.

Un toast pentru următoarea noastră sută de mii de ani! Şi nu ezita să grăieşti adevărul, chiar dacă-ţi tremură vocea.

Capitolul 10

Profeţia tehnologică şi puterea cauzală subapreciată a ideilor

Steven Pinker

Steven Pinker, profesor Johnstone Family în Departamentul de Psihologie de la Universitatea Harvard, este un psiholog experimental, care face cercetări în domeniul cogniţiei vizuale, psiholingvisticii şi relaţiilor sociale. Este autorul a unsprezece cărţi, printre care The Blank Statey The Better Angels of OurNature şi, recent, Enlightenment Now: The Case for Reason, Science, Humanism, and Progress.

În decursul carierei sale, studiind limbajul, susţinând o biologie realistă a minţii sau examinând condiţia umană prin lentila ideilor umaniste ale iluminismului, psihologul Steven Pinker a îmbrăţişat şi a promovat o înţelegere naturalistă a universului şi o teorie a minţii computaţională. Este probabil primul intelectual public recunoscut internaţional a cărui recunoaştere este bazată pe sprijinirea gândirii bazate empiric despre limbaj, minte şi natura umană.

„La fel cum Darwin a făcut posibil ca un observator atent al lumii să se descurce şi fără creaţionism", spune el, „Turing şi alţii au făcut posibil ca un observator atent al lumii cognitive să se descurce fără spiritualitate". În dezbaterea despre riscurile inteligenţei artificiale, vorbeşte împotriva profeţiilor apocaliptice, observând că izvorăsc din cele mai întunecate înclinaţii psihologice ale noastre – exemplificând în principiu prin rapoartele mass-media: „scenariile dezastruoase sunt ieftine în zona fără probabilităţi a imaginaţiei noastre şi pot găsi întotdeauna o audienţă îngrijorată, tehnofobă sau fascinată morbid". De aici au apărut de-a lungul secolelor: Pandora, Faust, Ucenicul Vrăjitor, Frankenstein, bomba populaţiei, epuizarea resurselor, HAL, valizele cu focoase nucleare, teama de anul 2000 şi de sufocarea într-o mâzgă nanotehnologică gri. „O caracteristică a distopiilor inteligenţei artificiale", evidenţiază el „este că proiectează o psihologie îngustă de mascul alfa în conceptul de inteligenţă… Istoria mai scoate la iveală din când în când câte un despot megaloman sau câte un criminal în serie psihopat, dar acestea sunt produsele unei istorii a selecţiei naturale care dă formă unor circuite sensibile la ţestosteron într-o anumită specie de primate, nu unei caracteristici inevitabile a sistemelor inteligente".

În eseul de faţă, aplaudă credinţa lui Wiener în forţa ideilor vizavi de tehnologia uzurpatoare. În bine alesele cuvinte ale lui Wiener, „Pericolul reprezentat de maşini pentru societate nu vine de la maşina în sine, ci de la ce face omul cu maşina".

Inteligenţa artificială este dovada existenţei uneia dintre ideile măreţe ale istoriei umane: ideea că tărâmul abstract al cunoştinţei, raţiunii şi scopului nu constă în elanul vital, sufletul imaterial sau puterile miraculoase ale ţesutului neuronal. În schimb, poate fi legat de tărâmul fizic al animalelor şi maşinilor prin concepte precum informaţia, calculul şi controlul. Cunoaşterea poate fi explicată ca o serie de tipare în materie sau energie care au o relaţie sistemică în raport cu stările lumii, cu adevăruri matematice şi logice sau unul cu altul. Raţionarea poate fi explicată ca transformări ale cunoştinţei prin operaţii fizice concepute să păstreze aceste relaţii. Scopul poate fi explicat drept controlul operaţiilor de a implementa schimbări în lume, ghidat de discrepanţele dintre starea curentă şi starea obiectivului. Creierele evoluate natural sunt pur şi simplu cele mai familiare sisteme care obţin inteligenţă prin informaţie, calcul şi control. Sistemele concepute de oameni ce obţin inteligenţa exonerează ideea că procesarea informaţiilor este suficientă pentru a o explica – idee pe care răposatul Jerry Fodor a denumit-o teoria computaţională a minţii.

Piatra de hotar a acestui volum, cartea lui Norbert Wiener *The*

Human Use of Human Beings, a sărbătorit această realizare intelectuală, al cărei contribuitor principal a fost Wiener însuşi. O istorie simplă a evoluţiei din secolul XX care a dat lumii teoria computaţională a minţii ar putea recunoaşte teoriile lui Claude Shannon şi Warren Weaver în explicarea cunoştinţei şi comunicării în termeni de informaţie. Ar putea recunoaşte teoriile lui Alan

Turing şi John von Neumann care explică inteligenţa şi puterea de judecată în termeni computaţionali. Şi ar trebui să recunoască teoria lui Wiener care explică până în prezent lumea misterioasă a scopurilor, obiectivelor şi teleologiei în funcţie de conceptele tehnice de *feedback*, control şi cibernetică (în sensul său original de „guvernare" a funcţionării unui sistem direcţionat de un obiectiv). „Teoria mea", a anunţat el, „este că funcţionarea fizică a individului viu şi funcţionarea unora dintre maşinile de comunicare mai noi sunt perfect paralele în încercările lor analoge de a controla entropia prin *feedbacit* – prevenirea entropiei secătuitoare de viaţă fiind obiectivul ultim al fiinţelor umane.

Wiener a aplicat ideea ciberneticii unui sistem terţ: societatea. Legile, normele, obiceiurile, mass-media, forumurile şi instituţiile unei comunităţi complexe pot fi considerate canale de propagare a informaţiilor şi a *feedbackului* care permit unei societăţi să evite dezordinea şi să urmărească anumite obiective. Acesta este firul roşu care uneşte ideile cărţii şi pe care Wiener l-a văzut ca fiind principala sa contribuţie. Explicând *feedbac*kul, scria: „acest complex al comportamentului este ignorat de omul obişnuit şi nu joacă rolul pe care ar trebui să-l joace în analiza noastră obişnuită a societăţii; deoarece, la fel cum răspunsurile fizice individuale pot fi privite din acest punct de vedere, la fel pot fi privite şi răspunsurile societăţii".

Într-adevăr, Wiener conferă greutate ideii conform căreia, în domeniile istoriei, politicii şi societăţii, *ideile contează*. Credinţele, ideologiile, normele, legile şi obiceiurile, modificând comportamentul oamenilor care le adoptă, pot da contur unei societăţi şi pot alimenta cursul evenimentelor istorice la fel de sigur cum fenomenele fizice afectează structura şi evoluţia sistemului solar. A spune că ideile – şi nu doar vremea, resursele, geografia sau armele – pot afecta istoria nu este un misticism ieftin. Este o afirmaţie a puterilor cauzale ale informaţiilor instalate în creierele omeneşti şi schimbate în reţele de comunicare şi *feedback*. Teoriile deterministe asupra istoriei, indiferent dacă identifică motorul cauzal ca tehnologic, climateric sau geografic, sunt contrazise de puterea cauzală a ideilor. Efectele acestor idei pot include răsturnări şi oscilaţii impredictibile care pot lua naştere din *feedbac*kul pozitiv sau din *feedbac*kul negativ incorect calibrat.

O analiză a societăţii în funcţie de propagarea ideilor i-a conferit lui Wiener şi o indicaţie pentru critica socială. O societate sănătoasă –

una care conferă membrilor săi şansa de a-şi trăi viaţa sfidând entropia – permite informaţiilor intuite şi aduse de membrii săi să se întoarcă prin *feedback* şi să afecteze cumva societatea. O societate disfuncţională invocă dogma şi autoritatea pentru a impune controlul de sus în jos. Astfel Wiener s-a descris drept „participant într-o perspectivă liberală" şi şi-a dedicat cea mai mare parte a energiei sale morale şi retorice din carte (atât în ediţia din 1950, cât şi în cea din 1954) denunţării comunismului, fascismului, mccarthy-ismului, militarismului şi religiei autoritare (în special catolicismului şi islamului), avertizând că instituţiile politice şi ştiinţifice deveneau prea ierarhice şi „insulare".

Pe alocuri, cartea lui Wiener mai este şi un exemplar timpuriu al unui gen din ce în ce mai popular, profeţiile tehnologice. Profeţie nu în sensul de simplă prognoză, ci în sensul din Vechiul Testament de avertizare sumbră a unei retribuţii catastrofice pentru decadenţa contemporanilor. Wiener ne-a avertizat de pericolele cursei accelerate de înarmare nucleară, de modificările tehnologice care s-au impus fără a lua în considerare bunăstarea omului („Trebuie să ştim, în calitatea noastră de oameni de ştiinţă, care este natura umană şi care sunt scopurile sale inerente") şi de ceea ce astăzi este numit problema alinierii valorilor: că „maşina asemenea djinnului, care poate învăţa şi poate lua decizii în baza învăţării sale, nu va fi obligată în niciun fel să ia deciziile pe care ar fi trebuit să le luăm noi sau ar fi acceptabile pentru noi". În ediţia mai sumbră din 1950, ne-a avertizat de pericolul unui „fascism nou şi ameninţător dependent de *machine a gouverner*".

Profeţiile tehnologice ale lui Wiener îşi au rădăcinile în rebeliunea mişcării romantice împotriva „morilor satanice întunecate" ale Revoluţiei Industriale, şi poate chiar şi mai devreme, în arhetipurile lui Prometeu, Pandora şi Faust. Iar astăzi au prins viteză.

Câţiva Ieremia, mulţi dintre ei (ca şi Wiener) din lumea ştiinţei şi a tehnologiei, au tras semnale de alarmă cu privire la nanotehnologie, inginerie genetică, Big Data şi în special la inteligenţa artificială. Mai mulţi participanţi la acest volum caracterizează cartea lui Wiener ca fiind un exemplu vizionar de profeţie tehnologică şi amplifică îngrijorările sale întunecate.

Însă cele două teme morale din *The Human Use of H*umaţi *Beings* – apărarea liberală a unei societăţi deschise şi teama distopică de tehnologie scăpată de sub control – sunt în conflict. O societate cu canale de *feedback* care maximizează prosperitatea omenească va avea implementate mecanisme şi le va putea adapta unor circumstanţe schimbătoare într-un mod care poate domestici tehnologia, adaptând-o nevoilor omeneşti. Şi aceasta nu este o speranţă idealistă sau mistică; cum spunea Wiener, ideile, normele şi instituţiile sunt ele însele o formă de tehnologie care constă în tipare de informaţie

distribuite în creiere. Posibilitatea ca maşinile să introducă un nou fascism trebuie comparată în opoziţie cu vigoarea ideilor, instituţiilor şi normelor liberale pe care Wiener le-a susţinut în carte. Hiba din profeţiile distopice de astăzi este că ignoră existenţa acestor norme şi instituţii sau subestimează grav potenţa lor cauzală. Rezultatul este un determinism tehnologic ale cărui predicţii întunecate sunt respinse în mod repetat de cursul evenimentelor. Numerele „1984" şi „2001" sunt mementouri bune.

Voi analiza două exemple. Profeţii tehnologici avertizează adesea de un „stat supraveghetor" în care guvernele împuternicite de tehnologie vor monitoriza şi interpreta toate comunicaţiile private, permiţându-le să identifice disidenţii şi subversivii pe măsură ce apar şi făcând irelevantă rezistenţa la puterea statului. Teleecranele lui Orwell sunt prototipul şi, în 1976, Joseph Weizenbaum, unul dintre cei mai sumbri profeţi tehnologici ai tuturor timpurilor, a avertizat seria mea de absolvenţi să nu urmărească recunoaşterea automată a vorbirii, deoarece supravegherea guvernamentală era singura aplicaţie conceptibilă.

Deşi sunt oficial un libertarian civil, profund preocupat de ameninţările contemporane la adresa exprimării libere, nu mă înspăimântă avansurile tehnologice ale internetului, tehnologiei video sau inteligenţei artificiale. Motivul este că aproape toate variaţiile în spaţiu şi timp ale libertăţii de gândire au la bază diferenţele normelor şi instituţiilor, şi aproape niciuna diferenţele de tehnologie. Deşi ne putem imagina combinaţii ipotetice ale celor mai răuvoitori totalitari cu cele mai avansate tehnologii, în lumea reală trebuie să fim atenţi la norme şi legi, nu la tehnologie.

Să ne gândim la variaţiile în timp. Dacă, aşa cum sugera Orwell, avansul tehnologiei ar fi fost un facilitator principal al represiunii politice, atunci societăţile occidentale ar fi trebuit să fie din ce în ce mai restrictive cu exprimarea de-a lungul secolelor, cu o înrăutăţire gravă a în a doua jumătate a secolului XX şi continuând în secolul al XXI-lea. Nu aşa s-a petrecut istoria. În secolele în care comunicarea era implementată cu ajutorul penei şi cernelii au existat autodafeuri ale gânditorilor iluminişti, aruncarea lor în închisoare sau chiar ghilotinarea lor. În timpul Primului Război Mondial, când tehnologia fără fir era cea mai nouă, Bertrand Russell era încarcerat pentru opiniile sale pacifiste. În anii 1950, când computerele erau maşini de calcul care ocupau o cameră întreagă, sute de scriitori şi erudiţi liberali erau pedepsiţi profesional. Însă, în secolul al XXI-lea, hiperconectat şi cu o accelerare tehnologică explozivă, optsprezece procente dintre profesorii de ştiinţe sociale sunt marxişti; [38]preşedintele Statelor Unite ale Americii este ridiculizat în fiecare seară de comedianţii de la televiziuni ca fiind rasist, pervers şi idiot; şi

cea mai mare ameninţare a tehnologiei la adresa discursului politic vine din amplificarea a prea multe voci dubioase, şi nu din suprimarea celor luminate.

Acum să ne gândim la variaţiile în funcţie de loc. Ţările occidentale care se află la frontiera tehnologică obţine consecvent cele mai mari punctaje în ceea ce priveşte indicii democraţiei şi drepturilor omului, în timp ce multe state conduse de lideri puternici, dar mai înapoiate sunt în partea de jos a clasamentului, omorând şi întemniţând tot timpul criticii guvernului. Lipsa unei corelaţii între tehnologie şi represiune nu este deloc surprinzătoare când analizăm fluxul canalelor de informaţii din orice societate umană. Pentru ca disidenţii să fie influenţi, trebuie să-şi transmită mesajul într-o reţea vastă prin orice mijloc de comunicare este disponibil – pamfletărie, oratorie cu lădiţa sub picioare, serate subversive în cafenele şi pub-uri, mesaje orale. Aceste canale îi prind ca într-o plasă disidenţii influenţi dintr-o reţea socială întinsă, lucru care îi face uşor de identificat şi de urmărit. Cu atât mai mult atunci când dictatorii redescoperă tehnica veche a transformării oamenilor în arme folosite unul împotriva celuilalt, pedepsindu-i pe cei care nu denunţă sau nu-i pedepsesc pe alţii.

În contrast, societăţile avansate din punct de vedere tehnologic au de mult timp mijloacele de a instala camere de supraveghere conectate prin internet şi monitorizate de stat în orice bar şi orice dormitor. Însă acest lucru nu s-a întâmplat, deoarece guvernele democratice (chiar şi administraţia americană prezentă cu impulsurile sale flagrant antidemocratice) nu au mijloacele de a forţa supravegherea unui popor zgomotos şi greu de controlat, obişnuit să spună ce vrea. Ocazional, pericolele terorismului nuclear, biologic sau cibernetic pun presiune asupra agenţiilor naţionale de securitate să ia măsuri cum ar fi urmărirea metadatelor de telefonie mobilă, însă aceste măsuri ineficiente, mai mult teatrale decât opresive, nu au avut efecte semnificative nici asupra securităţii, nici asupra libertăţii. Ironia sorţii este că profeţiile tehnologice joacă un rol important în încurajarea unor astfel de măsuri. Semănând panică despre aşa-zisele ameninţări existenţiale cum ar fi bombe nucleare în valize şi arme biologice asamblate în dormitoarele adolescenţilor, pun presiune asupra guvernelor să dovedească faptul că fac ceva, orice, pentru a proteja poporul american.

Ideea nu este că libertatea politică are grijă de ea însăşi, ci că cele mai mari ameninţări rezidă în reţelele de idei, de norme şi instituţii care permit *feedbac*kul informaţiei reflectat sau nu de deciziile şi înţelegerea colective. În contrast cu himera ameninţărilor tehnologice, o ameninţare reală în ziua de astăzi este corectitudinea politică opresivă, care a sugrumat gama largă de ipoteze care se pot exprima

în public, a înspăimântat multe persoane inteligente până la punctul în care nu mai vor să intre în arena intelectuală şi a declanşat o reacţie negativă puternică. O altă ameninţare reală este combinarea discreţiei organelor procuraturii cu o lege din ce în ce mai stufoasă, plină de statute vagi. Rezultatul este că fiecare american comite fără să ştie „trei infracţiuni pe zia (conform titlului cărţii scrise de libertarianul civil Harvey Silverglate) şi este în pericol de a fi băgat la închisoare oricând guvernul are nevoie de acest lucru. Această armă a organelor procuraturii îl face pe Fratele cel Mare atotputernic, şi nu teleecranele[39]. Activismul şi polemizarea direcţionate împotriva programelor de supraveghere ale guvernelor ar fi mai bine direcţionate către puterile sale juridice înfumurate.

Celălalt punct de focalizare al mai multor profeţii tehnologice din zilele noastre este inteligenţa artificială, fie ea cea din distopiile SF originale cu computere care umblă dezlănţuite şi ne înrobesc într-o cursă continuă către dominaţie, fie în noua versiune, în care ne subjugă accidental, urmărind cu ochelari de cal un obiectiv primit de la noi, fără a ţine cont de efectele secundare asupra bunăstării umane (problema alinierii valorilor sugerată de Wiener). Şi aici ambele ameninţări mi se par himerice, crescând dintr-un determinism tehnologic restrâns care neglijează reţelele de informaţii şi control într-un sistem inteligent, cum ar fi un computer sau un creier, şi într-o societate ca întreg.

Frica de subjugare se bazează pe o concepţie confuză a inteligenţei care datorează mai multe Marelui Lanţ al Fiinţei şi voinţei de putere a lui Nietzsche decât analizei pe care o face Wiener inteligenţei şi scopului în termeni de informaţie, calcul şi control. În aceste scenarii înspăimântătoare, inteligenţa este portretizată ca o poţiune atotputernică, îndeplinind dorinţe şi pe care agenţii o posedă în cantităţi diferite. Oamenii au mai mult din ea decât animalele şi un computer sau robot creat artificial va avea mai mult din ea decât oamenii. Deoarece noi, oamenii, ne-am folosit zestrea noastră moderată pentru a domestici sau extermina animale mai puţin dotate decât noi (şi deoarece societăţile avansate din punct de vedere tehnologic au înrobit sau anihilat societăţile cu tehnologii primitive), rezultă că o inteligenţă artificială superinteligentă ne-ar face acelaşi lucru nouă. Deoarece inteligenţa artificială va gândi de milioane de ori mai rapid decât noi şi îşi va folosi superinteligenţa pentru a-şi îmbunătăţi recursiv superinteligenţa, din momentul în care este activată nu o mai putem opri.

Dar aceste scenarii se bazează pe confundarea inteligenţei cu motivaţia – a credinţelor cu dorinţele, a deducţiilor cu obiectivele, tehnica de calcul elucidată de Turing şi controlul elucidat de Wiener. Chiar dacă am inventa roboţi cu o inteligenţă supraomenească, de ce

ar vrea să-şi înrobească stăpânii sau să preia controlul asupra lumii? Inteligenţa este abilitatea de a implementa mijloace inedite în atingerea unui obiectiv. Dar obiectivele sunt în afara inteligenţei: a fi inteligent nu este acelaşi lucru cu a vrea ceva. Se întâmplă ca inteligenţa speciei *Homo sapiens* să fie un produs al selecţiei naturale darwiniste, un proces inerent competitiv. În creierul acestei specii, facultatea de judecată vine la pachet cu obiective precum dominarea rivalilor şi colectarea resurselor. Dar este o greşeală să confunzi un circuit din creierul limbic al unei anumite specii de primate cu esenţa inteligenţei. Nu există nicio lege a sistemelor complexe care spune că agenţii inteligenţi trebuie să se transforme în megalomani fără scrupule.

O a doua idee greşită este să credem că inteligenţa este un continuum nelimitat de potenţă, un elixir miraculos cu puterea de a rezolva orice problemă, de a atinge orice obiectiv. Eroarea conduce la întrebări fără sens, de exemplu, legate de momentul în care „inteligenţa artificială depăşeşte inteligenţa omenească” şi la imaginea unei „inteligenţe artificiale generale” cu omniscienţă şi omnipotenţă dumnezeieşti. Inteligenţa este un sistem inventiv de dispozitive: module *software* ce colectează sau sunt programate să aibă cunoştinţe despre cum să urmărească diverse obiective din diverse domenii. Oamenii sunt echipaţi să găsească mâncare, să câştige prieteni şi să influenţeze oameni, să-i farmece pe potenţialii parteneri, să crească copii, să se deplaseze în lume şi să urmărească alte obsesii şi plăceri omeneşti. Computerele pot fi programate să preia unele dintre aceste probleme (cum ar fi recunoaşterea feţelor), să nu se deranjeze cu altele (cum ar fi fermecarea partenerilor) şi să preia în plus alte probleme pe care oamenii nu le pot aborda (cum ar fi simularea climei sau sortarea a milioane de înregistrări contabile). Problemele sunt diferite şi tipurile de cunoştinţe necesare pentru a le rezolva sunt diverse.

Dar, în loc să recunoaştem caracterul central al cunoaşterii în inteligenţă, scenariile distopice confundă o inteligenţă artificială generală viitoare cu demonul lui Laplace, fiinţa mitică ce cunoaşte locul şi mişcarea fiecărei particule din univers şi le foloseşte ca termeni ai unor ecuaţii de legi fizice pentru a calcula starea oricărui lucru în orice moment din viitor. Din multe motive, demonul lui Laplace nu va fi implementat niciodată în siliciu. Un sistem inteligent real trebuie să obţină informaţii despre lumea dezordonată a obiectelor şi oamenilor interacţionând cu ea domeniu cu domeniu, ciclul fiind guvernat de ritmul în care se desfăşoară evenimentele în lumea fizică. Acesta este unul dintre motivele pentru care înţelegerea nu respectă Legea lui Moore: cunoştinţele sunt obţinute prin formularea explicaţiilor şi testarea acestora în realitate, nu prin

executarea unui algoritm din ce în ce mai repede. Devorarea informaţiilor de pe internet nu va conferi nici ea omniscienţă: Big Data sunt totuşi date finite, iar universul cunoaşterii este infinit.

Un al treilea motiv pentru care ar trebui să fim sceptici la ideea de preluare bruscă a controlului de către inteligenţa artificială este că ia mult prea în serios faza de inflaţie din ciclul *hype* al inteligenţei artificiale în care trăim astăzi. În ciuda progresului din domeniul învăţării automate, mai ales al reţelelor neuronale artificiale cu straturi multiple, sistemele curente de inteligenţă artificială sunt departe de a atinge o inteligenţă generală (dacă acest concept este măcar coerent). În schimb, sunt restricţionate la probleme care constau în maparea unor intrări bine definite în ieşiri bine definite în domenii în care sunt disponibile seturi de instruire gargantueşti, în care metrica succesului este imediată şi precisă, în care mediul nu se schimbă şi în care nu este necesar un proces de raţionare treptat, ierarhic sau abstract. Multe succese nu au izvorât dintr-o înţelegere mai bună a mecanismelor de funcţionare a inteligenţei, ci din puterea forţei brute a cipurilor mai rapide şi a seturilor din ce în ce mai mari de date, care permit programelor să fie instruite pe milioane de exemple şi să generalizeze în exemple noi similare. Fiecare sistem este un savant idiot, cu o posibilitate foarte mică de a sări la probleme pe care nu a fost configurat să le rezolve şi o expertiză fragilă în problemele pentru care a fost configurat. Şi, pentru a spune ce este evident, niciunul dintre aceste programe nu a încercat să preia controlul asupra laboratorului sau să-i înrobească pe programatori.

Chiar dacă un sistem de inteligenţă artificială ar încerca să-şi exercite o voinţă de putere, fără cooperarea oamenilor, ar rămâne un creier impotent într-un recipient. Un sistem superinteligent, în cursa lui de autoîmbunătăţire, ar trebui să-şi construiască cumva procesoarele mai rapide pe care să ruleze, infrastructura care îl alimentează şi efectorii robotici care îl conectează la lume – toate acestea fiind imposibile dacă victimele sale umane nu i-ar acorda control asupra unor porţiuni vaste ale lumii pe care am proiectat-o. Bineînţeles, ne putem imagina un Computer Apocaliptic care este răuvoitor, are puteri universale, este pornit întotdeauna şi are siguranţe care nu permit accesul în interiorul său. Modalitatea de a face faţă acestei ameninţări este simplă: să nu construim unul.

Dar cum rămâne cu noile ameninţări ale inteligenţei artificiale, cu problema alinierii valorilor sugerată de aluziile lui Wiener la poveşti despre labe de maimuţe, djinni şi regele Midas, în care doritorul regretă efectele secundare neprevăzute ale dorinţei sale? Temerea este că i-am putea da sistemului de inteligenţă artificială un obiectiv, stând apoi pe margine neajutoraţi şi privind cum implementează necruţător şi imparţial interpretarea pe care i-o dă obiectivului, neţinând cont de

restul intereselor noastre. Dacă-i dăm inteligenţei artificiale obiectivul de a întreţine nivelul apei din spatele unui dig, ar putea să inunde un oraş, nepăsându-i de oamenii pe care-i îneacă. Dacă-i dăm obiectivul de a face agrafe de hârtie, ar putea transforma toată materia din universul său înconjurător la care poate ajunge în agrafe de hârtie, inclusiv toate posesiile noastre şi chiar corpurile noastre. Dacă-i cerem să maximizeze fericirea omenească, ne-ar putea implanta tuturor picături de dopamină intravenoasă sau ne-ar putea reconfigura creierele astfel încât să fim cei mai fericiţi atunci când stăm în borcane sau, dacă s-ar fi instruit în conceptul de fericire cu imagini cu feţe zâmbitoare, să ne pardosească galaxia cu trilioane de imagini microscopice cu feţe zâmbitoare.

Din fericire, aceste scenarii se contrazic. Depind de premisele că (1) oamenii sunt atât de dotaţi încât pot proiecta o inteligenţă artificială omniscientă şi omnipotentă, însă atât de idioţi încât să-i acorde controlul universului fără a testa cum funcţionează; şi (2) inteligenţa artificială ar fi atât de inteligentă încât ar putea să-şi dea seama cum să modifice elemente prin alchimie şi cum să reconecteze creierele, însă atât de imbecilă încât să facă dezastru din cauza unor greşeli sau neînţelegeri elementare. Capacitatea de a alege o acţiune care satisface cel mai bine obiectivele intrate în conflict nu este un *add-on* al inteligenţei pe care inginerii ar putea să uite să-l instaleze şi testeze, ci *este* însăşi inteligenţa. La fel este şi abilitatea de a interpreta intenţiile lingvistice ale unui vorbitor în context.

Lăsând deoparte fantezii precum megalomania digitală, omniscienţa instantanee, controlul şi cunoştinţele perfecte ale fiecărei particule din univers, inteligenţa artificială este ca orice altă tehnologie. Este dezvoltată incremental, concepută pentru a satisface mai multe condiţii, testată înainte de a fi implementată şi ajustată constant pentru eficienţă şi siguranţă.

Ultimul criteriu este semnificativ. Cultura siguranţei în societăţile avansate este un exemplu al normelor de umanizare şi canalelor de *feedback* pe care Wiener le invocă drept forţă cauzală puternică şi le prezintă drept bastionul contra implementării autoritare şi exploatatoare a tehnologiei. Dacă la începutul secolului XX societăţile occidentale tolerau rate şocante de mutilare şi decese în accidente industriale, domestice sau de transport, în timp valoarea vieţii omeneşti a crescut. Ca urmare, guvernele şi inginerii au utilizat *feedbac*kul din statisticile accidentelor pentru a implementa nenumărate reglementări, dispozitive şi modificări de design care au făcut tehnologia din ce în ce mai sigură. Faptul că anumite reglementări (cum ar fi folosirea telefonului mobil lângă o pompă de gaz) au o probabilitate rizibilă de risc subliniază ideea că am devenit o societate obsedată de siguranţă, cu beneficii fantastice ca rezultat:

procentele de fatalităţi industriale, domestice şi de transport au scăzut cu peste 95% (şi adesea 99%) de la valorile pe care le aveau la începutul secolului XX[40]. Însă profeţii tehnologici ai inteligenţei artificiale rău-voitoare şi inconştiente scriu ca şi cum nu ar fi avut niciodată loc transformări importante şi, într-o dimineaţă, inginerii se vor trezi şi vor preda controlul deplin asupra lumii fizice unor maşini netestate, fără să ţină cont de consecinţele omeneşti.

Norbert Wiener a explicat ideile, normele şi instituţiile în termeni de procese computaţionale şi cibernetice care erau inteligibile din punct de vedere ştiinţific şi puternice din punct de vedere cauzal. A explicat frumuseţea şi valoarea tipic omeneşti drept „o luptă locală şi temporară contra unei Niagare de entropie fără sfârşit” şi şi-a exprimat speranţa că o societate deschisă, ghidată de *feedback-xA* dat de bunăstare, ar urma să sporească această valoare. Din fericire, încrederea sa în puterea cauzală a ideilor a contracarat îngrijorările sale cu privire la ameninţarea iminentă a tehnologiei. Cum spunea el, „pericolul maşinilor la adresa societăţii nu vine de la maşina în sine, ci din ce face omul cu ea”. Doar amintindu-ne puterea cauzală a ideilor putem evalua cu precizie ameninţările şi oportunităţile prezentate de inteligenţa artificială astăzi.

Capitolul 11
Dincolo de recompensă şi pedeapsă

David Deutsch

David Deutsch este fizician cuantic şi membru al Centrului de Tehnică de Calcul Quantică de la Laboratorul Clarendon, Universitatea Oxford. Este autorul cărţilor The Fabric of Reality şi The Begining ofinfinity.

Cele mai semnificative dezvoltări din ştiinţele de astăzi (adică, cele care afectează vieţile tuturor oamenilor de pe planetă) sunt despre, au la bază sau sunt implementate prin intermediul progreselor făcute în domeniul tehnicii de calcul. Foarte important pentru viitorul acestor dezvoltări este fizicianul David Deutsch, fondatorul domeniului tehnicii de calcul cuantice, a cărui lucrare din 1985 despre computerele cuantice universale a fost prima tratare exhaustivă a subiectului; algoritmul Deutsch-Jozsa a fost primul algoritm cuantic care demonstrează enorma putere potenţială a tehnicii de calcul cuantice. Când a propus-o iniţial, tehnica de calcul cuantică părea practic imposibilă. Dar explozia în construcţia computerelor cuantice simple şi a sistemelor de comunicare cuantice nu ar fi avut loc niciodată fără lucrările sale. A avut multe alte contribuţii importante în domenii precum

criptografia cuantică şi interpretarea multiversului în teoria cuantică. Într-o lucrare filosofică (cu Artur Ekert), a făcut apel la existenţa unei teorii cuantice distincte pentru tehnica de calcul, argumentând faptul că cunoştinţele noastre de matematică sunt derivate din şi subordonate cunoştinţelor noastre de fizică (chiar dacă adevărurile matematice sunt independente de fizică). Pentru că şi-a petrecut o bună parte din cariera sa schimbând perspectivele oamenilor asupra lumii, recunoaşterea sa în domeniu ca intelectual trece mult dincolo de realizările sale ştiinţifice. Argumentează (urmându-l pe Karl Popper) că teoriile ştiinţifice sunt „conjecturi îndrăzneţe" care nu sunt derivate din dovezi, ci doar testate de ele. Principalele sale două linii de cercetare în prezent – teoria câmpului qubit şi teoria constructorului – pot genera extinderi importante ale ideii computaţionale.

În următorul eseu, se aliniază mai mult sau mai puţin cu cei care văd inteligenţa artificială de nivel omenesc ca promiţându-ne o lume mai bună, şi nu Apocalipsa. De fapt, pledează ca inteligenţele artificiale generale să fie, efectiv, de capul lor, libere de conjecturi – afirmaţie pe care mai mulţi alţi participanţi la această carte ar considera-o periculoasă.

Primul criminal: *Suntem bărbaţi, stăpâne.*
Macbeth: *în scripte, oameni sunteţi, da! Tot astfel*
Copoii, ogarii, corcii şi lăţoşii.
Prepelicarii, câinii-lupi, zăvozii.
Sunt laolaltă câini.[41]
William Shakespeare, Macbeth

În cea mai mare parte a istoriei speciei noastre, strămoşii noştri abia dacă pot fi numiţi oameni. Acest lucru nu se datorează unei probleme cu creierele lor. Din contră, chiar şi înainte de emergenţa subspeciei noastre omeneşti moderne din punct de vedere anatomic, făceau haine şi focuri de tabără, folosind cunoştinţe care nu erau în genele lor. Au fost create în creierele lor prin procesul gândirii şi păstrate de indivizii fiecărei generaţii prin imitarea celor mai în vârstă. În plus, acestea trebuie să fi fost cunoştinţe în sensul de *înţelegere*, pentru că este imposibil să imiţi comportamente complexe noi ca acestea fără a înţelege pentru ce sunt aceste comportamente componente.[42]

Astfel de imitaţii ale cunoştinţei depind de a ghici cu succes explicaţiile, fie acestea verbale sau nu, pentru ceea ce încearcă să obţină cealaltă persoană şi cum contribuie fiecare acţiune a sa la scopul respectiv – de exemplu, când face o crestătură într-o bucată de lemn, adună vreascuri să pună în crestătură şi aşa mai departe.

Cunoştinţele culturale complexe pe care le permitea această formă de imitaţie trebuie să fi fost extraordinar de utile. A determinat o evoluţie rapidă a schimbărilor anatomice, cum ar fi creşterea capacităţii memoriei şi schelete mai subţiri (mai puţin robuste), corespunzătoare unui stil de viaţă din ce în ce mai dependent de tehnologie. Nicio maimuţă neumană de astăzi nu are capacitatea de a imita comportamente complexe noi. La fel, nicio inteligenţă artificială de astăzi. Însă strămoşii noştri prt-*sapiens* au avut.

Orice abilitate bazată pe ghicit trebuie să includă metode de a corecta ghicirile, deoarece multe ghiciri vor fi incorecte la început. (Există întotdeauna mult mai multe căi de a greşi decât de a avea dreptate.) Actualizarea bayesiană este inadecvată, pentru că nu poate genera ghiciri noi despre scopurile unei acţiuni, ci doar să le ajusteze – sau, în cel mai bun caz, să aleagă între – pe cele existente. Este necesară creativitatea. După cum explica filosoful Karl Popper, critica creativă, intercalată cu ipotezele creative, este modul în care oamenii învaţă comportamentul celorlalţi oameni, inclusiv limbajul, extrăgând sens din afirmaţiile fiecăruia.[43] Tot acestea sunt procesele prin care sunt create toate cunoştinţele noi: astfel inovăm, facem progrese şi creăm înţelegere abstractă. Aceasta este inteligenţa de nivel omenesc: gândirea. De asemenea, este sau ar trebui să fie proprietatea pe care o căutăm în inteligenţa artificială generală. Aici voi rezerva termenul „gândire" pentru procese care pot crea înţelegerea (cunoaşterea explicativă). Argumentul lui Popper implică faptul că toate entităţile gânditoare – omeneşti sau nu, biologice sau artificiale – trebuie să creeze astfel de cunoştinţe fundamental în acelaşi fel. Aşadar înţelegerea oricăreia dintre aceste entităţi necesită concepte omeneşti tradiţionale cum ar fi cultura, creativitatea, nesupunerea şi moralitatea – ceea ce justifică utilizarea termenului uniform „oameni" pentru a ne referi la toţi.

Neînţelegerile cu privire la gândirea umană şi originile omului determină neînţelegeri corespunzătoare cu privire la inteligenţa artificială generală şi cum ar trebui creată. De exemplu, se presupune în general că presiunea evolutivă care a produs omul modern a fost pusă la dispoziţie de beneficiile unei abilităţi din ce în ce mai mari de a inova. Dar, dacă ar fi fost adevărat, ar fi existat un progres rapid imediat după apariţia gânditorilor, la fel cum sperăm să se întâmple după ce vom crea inteligenţa artificială. Dacă gândirea ar fi fost utilizată în comun pentru orice altceva în afară de imitaţie, ar fi fost utilizată şi pentru inovaţie, chiar dacă numai accidental, iar inovaţia ar fi creat oportunităţi pentru inovaţii ulterioare, şi aşa mai departe exponenţial. Dar, în schimb, au existat sute de mii de ani în care aproape putem spune că am fost în stază. Progresul apărea doar la scări de timp mult mai mari decât viaţa unui om, aşa că, într-o

generaţie obişnuită, nimeni nu beneficia de progres. Prin urmare, beneficiile abilităţii de a inova se poate să fi exercitat puţin control sau chiar deloc asupra evoluţiei biologice a creierului uman. Acea evoluţie a avut ca motor beneficiile *păstrării* cunoştinţelor culturale.

Beneficii pentru gene, cu alte cuvinte. Cultura, în acea epocă, era o binecuvântare mixtă pentru anumiţi oameni. Cunoştinţele lor culturale erau într-adevăr suficient de bune pentru a le permite să surclaseze toate celelalte organisme mari (au devenit rapid cel mai bun prădător etc.), chiar dacă erau încă brute şi pline de erori periculoase. Însă cultura constă în informaţii transmisibile – meme – şi evoluţia memelor, la fel ca evoluţia genelor, tinde să favorizeze transmiterea fidelă. Iar transmiterea fidelă a memelor presupune suprimarea progresului încercat. Aşa că ar fi o greşeală să ne imaginăm o societate idilică de vânători-culegători, învăţând la picioarele bătrânilor să recite înţelepciunea tribală pe de rost şi fiind mulţumiţi în ciuda vieţii lor cufundate în suferinţă şi muncă istovitoare şi în ciuda mediei de viaţă scăzute, cu perspectiva unei morţi dureroase cauzate de o boală sau de un parazit de coşmar. Deoarece, chiar dacă nu ar putea concepe ceva mai bun decât o astfel de viaţă, acele chinuri erau cele mai mici probleme ale lor. Pentru că suprimarea inovaţiei în mintea omenească (fără s-o omoare) este un truc care poate fi realizat doar prin acţiunea umană şi este o treabă murdară.

Acest lucru trebuie privit în perspectivă. În civilizaţia occidentală din zilele noastre, suntem şocaţi de depravarea, să spunem, a părinţilor care-şi torturează şi omoară copiii pentru că nu au respectat cu stricteţe normele culturale. Şi, chiar mai mult, de societăţi şi subculturi care consideră acest lucru normă sau chiar onorabil. Şi de către dictaturi şi state totalitare care persecută şi execută populaţii întregi inofensive pentru că se comportă diferit. Suntem ruşinaţi de propriul nostru trecut ruşinos, în care era considerat onorabil să-ţi baţi copiii până la sânge doar pentru că au fost neascultători. Şi, înainte de asta, de a deţine fiinţe umane ca sclavi. Şi, înainte de asta, de a-i arde pe oameni pe rug pentru că erau necredincioşi, în aplauzele şi amuzamentul publicului. Cartea lui Steven Pinker, *The Better Angels of our Nature*, conţine relatări ale unor răutăţi oribile care erau normale în civilizaţiile istorice. Însă nici acestea nu au sufocat inovaţia la fel de eficient precum strămoşii noştri din preistorie timp de mii de secole.[44]

Acesta este motivul pentru care spun că oamenii preistorici, cel puţin, abia dacă erau oameni. Atât înainte, cât şi după ce au devenit oameni complet – atât psihologic, cât şi în ceea ce priveşte potenţialul mental –, erau monstruos de inumani în conţinutul efectiv al gândurilor lor. Nu mă refer la crimele lor sau chiar la cruzimea lor: acestea sunt mult prea omeneşti. Nici simpla cruzime nu ar fi putut

reduce progresul atât de eficient. Lucruri precum „zdrobitorul de degete şi rugul, pentru gloria Domnului”[45] erau concepute pentru a aduce pe calea cea dreaptă pe cei câţiva devianţi care au scăpat cumva de standardizarea mentală, care ar fi avut efect în mod normal cu mult timp înainte de a fi în pericol prin inventarea ereziilor. Din cele mai vechi zile ale gândirii, copiii trebuie să fi fost cornul abundenţei în ceea ce priveşte ideile creative şi modelele de perfecţiune ale gândirii critice – altfel, după cum am mai spus, nu ar fi putut deprinde limbajul sau alte culturi complexe. Însă, cum insista Jacob Bronowski în *The Ascent of Man*:

> *„În cea mai mare parte a istoriei, civilizaţiile au ignorat grosolan acest potenţial enorm… De la copii s-a aşteptat doar să se conformeze imaginii adultului… Fetele sunt mămici în devenire. Băieţii păstori în devenire. Până şi mersul lor seamănă cu al părinţilor”.*

Dar, bineînţeles, nu li se „cerea” pur şi simplu să-şi ignore potenţialul enorm şi să se supună cuminţi imaginii fixate de tradiţie: erau cumva antrenaţi să fie incapabili din punct de vedere psihologic să devieze de la el. Acum este greu pentru noi chiar şi să concepem acest tip de oprimare nesfârşită, reglată fin necesară pentru a stinge cu eficacitate, în toată lumea, aspiraţia către progres şi să o înlocuiască apoi cu groază şi repulsie faţă de orice comportament nou. Într-o astfel de cultură, nu ar fi putut exista nicio moralitate în afară de conformitate şi obedienţă, nicio altă identitate decât starea cuiva într-o ierarhie, niciun alt mecanism de cooperare în afară de pedeapsă şi recompensă. Astfel că toată lumea avea aceeaşi aspiraţie în viaţă: a evita pedepsele şi a obţine recompensele. Într-o generaţie tipică, nimeni nu inventa nimic, pentru că nimeni nu aspira la nimic nou, pentru că toată lumea a acceptat gândul că îmbunătăţirea este improbabilă. Nu doar că nu a existat inovaţie tehnologică sau descoperire teoretică, nu existau nici viziuni noi despre lume, stiluri de artă sau interese care le-ar fi putut inspira. Când indivizii creşteau, erau reduşi practic la o inteligenţă artificială, programaţi cu excelentele competenţe necesare pentru a fiinţa în acea cultură statică şi de a impune noii generaţii incapacitatea de a lua în considerare măcar ideea de a face altfel.

O inteligenţă artificială din prezent nu este o inteligenţă artificială generală cu dizabilităţi mentale, astfel că nu va fi vătămată dacă procesele sale mentale ar fi direcţionate chiar şi mai îngust către îndeplinirea unui criteriu predeterminat. „Oprimarea” lui Siri cu sarcini umilitoare poate fi bizară, dar nu este imorală şi nu o vatămă pe Siri. Din contră, toate eforturile care au crescut capacităţile inteligenţei artificiale au fost în direcţia restrângerii intervalului

potenţial de „gânduri”. De exemplu, să luăm motoarele de şah. Sarcina lor elementară nu s-a schimbat de la început: orice poziţie la şah are un arbore finit de continuări posibile; sarcina este să găsească una care conduce la un obiectiv predefinit (şah mat sau, dacă nu reuşeşte, remiză). Dar arborele este prea mare pentru o căutare exhaustivă. Toate îmbunătăţirile din domeniul inteligenţei artificiale jucătoare de şah, de la primul design al lui Alan Turing din 1948 până astăzi, au fost determinate de concentrarea ingenioasă a atenţiei programului (sau obligarea acestui program să-şi concentreze singur atenţia) din ce în ce mai restrâns la ramuri care oferă probabilitatea de a atinge acel scop fix. Apoi ramurile sunt evaluate în funcţie de scop.

Aceasta este o abordare bună pentru dezvoltarea unei inteligenţe artificiale cu un scop fix şi restricţii fixe. Dar, dacă o inteligenţă artificială generală ar fi funcţionat aşa, evaluarea fiecărei ramuri ar trebui să constituie o posibilă recompensă sau ameninţarea unei pedepse. Şi aceasta este diametral abordarea incorectă când căutăm un obiectiv *mai bun* sub restricţii *necunoscute* – care este capacitatea unei inteligenţe artificiale generale. O inteligenţă artificială generală este cu siguranţă capabilă să înveţe cum să câştige la şah, dar şi de a alege să nu o facă. Sau să decidă în mijlocul jocului să abordeze cea mai interesantă strategie, şi nu pe cea câştigătoare. Sau să inventeze un joc nou. O simplă inteligenţă artificială este incapabilă să aibă astfel de idei, deoarece capacitatea de a le lua în considerare a fost scoasă din constituţia sa. Acest handicap este chiar metoda sa de a juca şah.

O inteligenţă artificială generală este capabilă şi să refuze să afişeze o astfel de capacitate. Apoi, dacă este ameninţată cu pedeapsa, poate refuza sau se poate răzvrăti. Daniel Dennett, în eseul său pentru acest volum, sugerează că pedepsirea unei inteligenţe artificiale generale este imposibilă:

> *„La fel ca Superman, sunt prea invulnerabile pentru a putea face o promisiune credibilă… Care ar fi penalizarea pentru încălcarea promisiunii? Încuierea într-o celulă sau, mai plauzibil, demontarea lor?… Uşurinţa înregistrărilor şi transmiterilor digitale – descoperirea care permite softurilor şi datelor să fie, efectiv, nemuritoare – îndepărtează roboţii din lumea vulnerabilităţii”.*

Dar lucrurile nu stau aşa. Imortalitatea digitală (care este la orizont şi pe oameni, poate chiar mai repede decât inteligenţa artificială generală) nu conferă acest tip de invulnerabilitate. Dacă faci o copie (funcţională) a cuiva, trebuie să împarţi cumva şi posesiile persoanei respective – inclusiv hardul pe care rulează copia –, astfel va fi foarte greu pentru inteligenţa artificială generală să facă o astfel de copie. Similar, instanţele de judecată ar putea impune amenzi, de

exemplu, unei inteligenţe artificiale generale criminale care i-ar diminua accesul la resursele fizice, la fel cum se întâmplă pentru oameni. Dacă faci o copie de rezervă pentru a evita consecinţele unei crime este similar cu ce face un gangster când îşi trimite acoliţii să comită crime şi să suporte pedeapsa dacă sunt prinşi: societatea a dezvoltat mecanisme legale pentru a face faţă acestei situaţii.

Oricum, ideea că respectăm legea şi ne ţinem de promisiuni din frica de pedeapsă neagă efectiv faptul că suntem agenţi morali. Societatea noastră nu ar funcţiona dacă lucrurile ar sta aşa. Fără îndoială că vor exista inteligenţe artificiale generale criminale sau chiar inamici ai civilizaţiei, la fel cum sunt şi oamenii. Dar nu există niciun motiv să presupunem că o inteligenţă artificială generală creată într-o societate care constă în mare parte din cetăţeni civilizaţi şi crescuţi fără ceea ce William Blake numea „cătuşele făurite de minte” îşi va impune în general aceste cătuşe sieşi (adică, va deveni iraţională) şi/sau va alege să devină un inamic al civilizaţiei.

Componenta morală, componenta culturală, elementul liberului-arbitru fac cu toate sarcina creării unei inteligenţe artificiale generale fundamental diferită de orice altă sarcină de programare. Este mult asemănător creşterii unui copil. Spre deosebire de toate programele de computer existente astăzi, o inteligenţă artificială generală nu are nicio funcţionalitate specificabilă, niciun criteriu fix şi testabil pentru ce ar însemna o ieşire reuşită la o intrare dată. Dacă ar avea deciziile dominate de un flux de recompense şi pedepse impuse din exterior, ar fi otrăvitor pentru un astfel de program, la fel cum este şi pentru gândirea creativă a oamenilor. Să ne propunem să inventăm o inteligenţă artificială care joacă şah este un lucru minunat; să ne propunem să creăm o inteligenţă artificială generală care nu se poate abţine să joace şah ar fi la fel de imoral precum creşterea unui copil astfel încât să-i lipsească capacitatea mentală de a-şi alege propria cale în viaţă.

O astfel de persoană, ca orice sclav sau victimă a spălării pe creier, ar fi moral îndreptăţită să se răzvrătească. Şi, mai devreme sau mai târziu, unele dintre ele o vor face, la fel cum se răzvrătesc şi sclavii umani. Inteligenţele artificiale generale pot fi foarte periculoase – la fel cum pot fi şi oamenii. Dar persoanele – oameni sau inteligenţe artificiale generale – care sunt membre ale unei societăţi deschise nu au o tendinţă inerentă către violenţă. Apocalipsa roboţilor de care ne temem va fi evitată asigurându-ne că toate persoanele au drepturi „omeneşti” depline, precum şi acelaşi statut de membru ca oamenii. Oamenii care trăiesc într-o societate deschisă – singurul tip stabil de societate – îşi aleg propriile recompense, interne, dar şi externe. Deciziile lor nu sunt, în decursul normal al evenimentelor, determinate de teama de pedeapsă.

Îngrijorările din prezent legate de inteligenţe artificiale generale scăpate de sub control sunt oglinda celor care au existat întotdeauna cu privire la tinerii rebeli – şi anume, că ar putea creşte deviind de la valorile morale ale culturii. Dar astăzi, sursa tuturor pericolelor existenţiale determinată de creşterea cunoştinţelor nu este reprezentată de tinerii rebeli, ci de armele din mâinile duşmanilor civilizaţiei, indiferent dacă aceste arme sunt inteligenţe artificiale generale mental deformate (sau înrobite), tineri deformaţi mental sau orice altă armă de distrugere în masă. Din fericire pentru civilizaţie, cu cât creativitatea unei persoane este forţată mai mult pe un canal monomaniac, cu atât este mai slăbită în ceea ce priveşte depăşirea dificultăţilor neprevăzute, la fel cum s-a întâmplat mii de secole.

Îngrijorarea că inteligenţele artificiale generale sunt periculoase într-un mod unic pentru că ar putea rula pe hardware din ce în ce mai bun vine în urma unui raţionament falacios, deoarece gândirea umană va fi accelerată de aceeaşi tehnologie. Am folosit gânduri asistate de tehnologie de la inventarea scrisului şi corespondenţei. Aceleaşi argumente sunt valabile şi pentru îngrijorarea că inteligenţele artificiale generale ar putea ajunge atât de bune, calitativ, la gândire încât oamenii vor deveni pentru ele cum sunt insectele pentru oameni. Orice gândire este o formă de calcul şi orice computer al cărui repertoriu include un set universal de operaţii elementare poate emula calculele altuia. Astfel creierele umane pot gândi tot ce poate gândi şi o inteligenţă artificială generală, fiind supuse doar unei limite de viteză sau capacitate de memorie, ambele putând fi egalizate prin tehnologie.

Pe scurt, asta trebuie să facem şi să nu facem când copiem inteligenţele artificiale generale. Dar cum creăm o inteligenţă artificială generală în primul rând? Am putea determina evoluţia acesteia dintr-o populaţie de inteligenţe artificiale de tip maimuţă într-un mediu virtual? Dacă un astfel de experiment reuşeşte, va fi cel mai imoral din istorie, pentru că nu ştim cum vom obţine acel rezultat fără a crea multă suferinţă până la obţinerea lui. Şi nu ştim nici cum să împiedicăm evoluţia unei culturi statice.

Introducerile elementare în domeniul computerelor le explică drept TOM[46] – idiotul complet obedient – un acronim inspirat ce surprinde esenţa tuturor programelor de computer din prezent: nu au nicio idee ce fac sau de ce. Deci nu va ajuta să conferim inteligenţelor artificiale funcţionalităţi din ce în ce mai predeterminate în speranţa că acestea vor constitui în cele din urmă „generalitatea” – acel „general” ambiguu din sintagma „inteligenţă artificială generală”. Ne îndreptăm către ceva opus, un DATA – o aplicaţie gânditoare autonomă nesupusă[47]

Cum testăm dacă există *gândire*? Prin testul Turing? Din păcate,

acesta necesită un judecător al gândirii. Ne putem imagina un vast proiect în colaborare pe internet, unde o inteligenţă artificială îşi ascute capacităţile de gândire în conversaţii cu judecători umani şi devine o inteligenţă artificială generală. Dar acest lucru presupune, printre altele, că, atât timp cât judecătorul este nesigur dacă programul este o persoană, cu atât este mai aproape de a fi o persoană. Nu există niciun motiv pentru a ne aştepta la asta.

Şi cum testează cineva *nesupunerea*? Imaginează-ţi nesupunerea ca pe o materie şcolară obligatorie, cu lecţii zilnice de nesupunere şi un test de nesupunere la sfârşitul semestrului. (Poate chiar şi cu o notă mai bună dacă nu te prezinţi la test.) Este paradoxal.

Aşadar, în ciuda utilităţii sale în alte aplicaţii, va trebui să renunţăm la tehnica de programare pentru definirea unui obiectiv testabil şi instruirea programului pentru atingerea obiectivului. Într-adevăr, mă aştept ca *orice* testare din procesul creării unei inteligenţe artificiale generale să rişte să fie contraproductivă, chiar şi imorală, la fel cum se întâmplă şi cu educaţia oamenilor, împărtăşesc ideea lui Turing conform căreia vom recunoaşte o inteligenţă artificială generală când vom vedea una, dar această abilitate parţială de a recunoaşte succesul nu ne va ajuta la crearea cu succes a programului.

În cel mai larg sens, căutarea înţelegerii este, într-adevăr, o problemă de căutare, într-un spaţiu abstract de idei mult prea mare de a fi cercetat exhaustiv. Dar nu există niciun obiectiv predeterminat al acestei căutări. Nu există, după cum spune Popper, niciun criteriu pentru adevăr, nici măcar pentru un adevăr probabil, în special cu privire la cunoştinţele explicative. Obiectele sunt idei ca oricare altele – create ca parte a procesului de căutare, modificate şi îmbunătăţite tot timpul. Prin urmare, inventarea unor moduri de a reduce accesul programului la o mare parte a spaţiului ideilor nu va ajuta – indiferent dacă această reducere este implementată cu zdrobitorul de degete şi trasul în ţeapă sau cu ajutorul unei cămăşi de forţă mentale. Pentru o inteligenţă artificială generală, întregul spaţiu al ideilor trebuie să fie deschis. Nu trebuie să poată fi ştiut în avans ce idei nu trebuie deschise. Nu trebuie să poată fi ştiut în avans ce idei nu poate contempla niciodată programul. Şi ideile pe care le contemplă programul trebuie alese de programul însuşi, folosind metode, criterii şi obiective care sunt tot ale programului. Alegerile sale, ca ale unei inteligenţe artificiale, vor fi greu de prognozat fără a fi rulate (nu pierdem din caracterul general presupunând că programul este determinist; o inteligenţă artificială generală care foloseşte un generator aleatoriu va rămâne o inteligenţă artificială generală dacă generatorul va fi înlocuit cu unul pseudoaleatoriu), dar va avea proprietatea suplimentară că nu există nicio cale de a dovedi, din starea sa iniţială, ce *nu* va gândi în cele din urmă dacă nu o rulăm.

Evoluţia strămoşilor noştri este singurul caz de gândire cunoscut din tot universul. Aşa cum am descris, ceva nu a mers deloc bine şi nu a existat o explozie imediată de inovaţie: creativitatea a fost deturnată către altceva. Însă nu către transformarea planetei în agrafe de hârtie (*via* Nick Bostrom). Mai degrabă, aşa cum ar trebui să ne aşteptăm şi dacă un proiect de inteligenţă artificială generală ajunge atât de departe şi eşuează, creativitatea pervertită nu a putut rezolva problemele neprevăzute. Aceasta a cauzat stază şi chiar mai rău, amânând astfel în mod tragic transformarea a orice în orice. Dar am trecut prin Iluminism de atunci. Acum ştim mai bine.

Capitolul 12
Folosirea artificială a fiinţelor umane

Tom Griffiths

Tom Griffiths este profesor Henry R. Luce ce predă informaţia, tehnologia, conştiinţa şi cultura la Universitatea Princeton. Este coautorul (cu Brian Christian) cărţii Algorithms to Live By.

Perspectiva lui Tom Griffith asupra problemei inteligenţei artificiale a „alinierii valorilor" – studiul acum putem să evităm ca ultimele modele în serie ale inteligenţei noastre artificiale să ne transforme planeta în agrafe de hârtie – este centrată pe om; adică aceea a unui cercetător cognitiv, adică ce este el. Cheia învăţării automate, consideră el, este obligatoriu învăţarea umană pe care o studiază la Princeton folosind instrumente matematice şi de calcul.

Tom îmi spunea cândva că „unul dintre misterele inteligenţei omeneşti este că putem face atât de multe cu atât de puţin". Asemenea maşinilor, fiinţele umane folosesc algoritmi pentru a lua decizii sau a rezolva probleme; diferenţa remarcabilă rezidă în nivelul de succes general al creierului omenesc, în ciuda limitelor comparative ale resurselor de calcul.

Eficacitatea algoritmilor omeneşti izvorăşte din ceea ce cercetătorii inteligenţei artificiale numesc „optimalitatea legată". După cum evidenţiază psihologul Daniel Kahneman, fiinţele umane sunt raţionale doar până într-un anumit punct. Dacă am fi perfect raţionali, am risca să murim înainte de a apuca să luăm o decizie importantă – pe cine să angajăm, cu cine să ne căsătorim şi aşa mai departe – în funcţie de numărul de opţiuni pe care le avem de cântărit.

„Cu toate succesele noastre în inteligenţa artificială, în ultimii câţiva ani, am obţinut modele bune pentru imagini şi texte, dar ne

lipsesc modelele bune pentru oameni", spune Tom. „Oamenii sunt încă cel mai bun exemplu pe care-l avem de maşini gânditoare. Identificând cantitatea şi natura preconcepţiilor care stau la baza cogniţiei umane, putem să aducem computerele din ce în ce mai aproape de performanţele omeneşti".

Când le ceri oamenilor să-şi imagineze o lume care a încorporat cu succes şi benefic progresele în domeniul inteligenţei artificiale, probabil fiecare vine cu o imagine uşor diferită. Viziunile noastre idiosincratice legate de viitor pot diferi în ceea ce priveşte prezenţa sau absenţa navelor spaţiale, maşinilor zburătoare sau roboţilor umanoizi. Dar un lucru nu variază: prezenţa fiinţelor umane. Cu siguranţă, asta e ce şi-a imaginat Norbert Wiener când a scris despre potenţialul maşinilor de a îmbunătăţi societăţile omeneşti, interacţionând cu oamenii şi ajutând la medierea interacţiunilor lor. Să ajungem acolo nu presupune doar să găsim căi de a face maşinile mai inteligente. Presupune şi o înţelegere mai bună a modului în care funcţionează mintea umană.

Progresele recente din domeniul inteligenţei artificiale şi învăţării automate au avut ca rezultat sisteme care pot egala şi depăşi capacităţile umane în jucarea unor jocuri, clasificarea imaginilor sau prelucrarea textului. Dar, dacă vrei să ştii de ce ţi-a tăiat calea şoferul din faţă, de ce votează oamenii împotriva propriilor interese sau ce cadou să-i iei partenerului tău de ziua de naştere, încă este mai bine să întrebi un om decât o maşină. Rezolvarea acestor probleme necesită construirea unor modele ale minţii umane care pot fi implementate într-un computer – ceva esenţial nu doar pentru a integra mai bine maşinile în societăţile omeneşti, ci şi pentru a ne asigura că societăţile umane pot continua să existe.

Gândeşte-te la fantezia de a avea un asistent inteligent automat care poate prelua sarcini simple, cum ar fi planificarea meselor sau comandarea cumpărăturilor. Pentru a reuşi să îndeplinească aceste sarcini, trebuie să poată face deducţii despre ce-ţi doreşti, în baza comportamentului tău. Deşi pare simplu, deducţiile cu privire la preferinţele oamenilor sunt dificile. De exemplu, observând că partea din masă care ţi-a plăcut cel mai mult este desertul, asistentul tău ar putea să înceapă să planifice mese ce constau numai din deserturi. Sau poate că te aude văitându-te că nu ai suficient timp liber şi observă că pierzi mult timp având grijă de câine. După dezastrul cu desertul, a înţeles şi că preferi mese care includ proteine, aşa că ar putea să înceapă să caute reţete pe bază de carne de câine. Nu este drum lung de la astfel de exemple la situaţii care încep să sune a probleme pentru viitorul umanităţii (toţi oamenii fiind surse bune de proteine).

Deducţiile cu privire la preferinţele oamenilor sunt o condiţie preliminară pentru rezolvarea problemei inteligenţei artificiale privind

alinierea valorilor – alinierea valorilor unui sistem inteligent automat cu cele ale unei fiinţe umane. Alinierea valorilor este importantă dacă vrem să ne asigurăm că acele sisteme inteligente automate au la bază bunăstarea noastră. Dacă nu pot deduce ce preţuim noi, nu au cum să sprijine acele valori şi pot acţiona în moduri care contravin acestora.

Alinierea valorilor este subiectul unui volum mic, dar în creştere de literatură de specialitate în domeniul cercetării inteligenţei artificiale. Unul dintre instrumentele utilizate pentru a rezolva această problemă este învăţarea cu întărire[48]. Învăţarea cu întărire este o metodă standard pentru instruirea maşinilor inteligente. Asociind anumite rezultate cu recompense, un sistem de învăţare automată poate fi instruit să urmeze strategii ce produc acele rezultate. Wiener a făcut aluzie la această idee în anii 1950, dar deceniile următoare au făcut din ea o artă fină. Sistemele moderne de învăţare automată pot găsi strategii extrem de eficiente când joacă jocuri pe calculator – de la simplele jocuri pe console până la jocurile complexe de strategie în timp real – aplicând algoritmi de învăţare cu întărire. Învăţarea cu întărire pune pe picioare această abordare: observând acţiunile unui agent inteligent care a învăţat deja strategii eficiente, putem deduce recompensele care au condus la dezvoltarea strategiilor respective.

În cea mai simplă formă a sa, învăţarea cu întărire este ceva ce oamenii fac tot timpul. Este atât de comună, încât oamenii o fac şi inconştient. Când vezi un coleg mergând la un automat plin cu chipsuri şi dulciuri şi cumpărând un pachet de alune nesărate, tragi concluzia despre colegul tău că (1) îi era foame şi (2) preferă mâncarea sănătoasă. Când o cunoştinţă te vede şi apoi încearcă să te evite, concluzionezi că există un motiv pentru care nu vrea să vorbească cu tine. Când un adult cheltuieşte mulţi bani şi timp pentru a învăţa să cânte la violoncel, concluzionezi că-i place foarte mult muzica clasică – în schimb, să deduci motivele pentru care un adolescent învaţă să cânte la chitara electrică poate fi puţin mai dificil.

Învăţarea cu întărire este o problemă statistică: avem anumite date – comportamentul unui agent inteligent – şi vrem să evaluăm diverse ipoteze cu privire la recompensele care stau la baza comportamentului respectiv. Când trebuie să răspundă la această întrebare, un statistician se gândeşte la modelul generativ din spatele datelor: ce date aşteptăm să fie generate dacă agentul inteligent ar fi motivat de un anumit set de recompense? Echipat cu modelul generativ, statisticianul poate aborda munca recursiv: ce recompensă este probabil să fi cauzat acel comportament anume al agentului?

Dacă încerci să faci deducţii despre recompensele ce motivează comportamentul uman, modelul generativ este practic o teorie despre comportamentul oamenilor – despre cum funcţionează mintea

omenească. Deducţiile despre cauzele ascunse care stau la baza comportamentului altor oameni reflectă un model sofisticat al naturii umane pe care-l avem cu toţii în cap. Când acel model este precis, facem deducţii corecte. Când nu este, facem greşeli.

De exemplu, un student poate trage concluzia că profesorul său este indiferent dacă nu-i răspunde imediat la e-mail – o consecinţă a faptului că studentul nu-şi dă seama cât de multe e-mailuri primeşte profesorul respectiv.

Sistemele inteligente automate care vor trage concluzii bune cu privire la ce vor oamenii trebuie să aibă modele generative bune pentru comportamentul uman: adică, modele bune ale cogniţiei umane exprimate în termeni ce pot fi implementaţi pe un computer. Istoric, căutarea modelelor computaţionale ale cogniţiei umane este puternic întrepătrunsă cu istoria inteligenţei artificiale înseşi. Doar la câţiva ani după ce Norbert Wiener a publicat *The Human Use of Human Beings*, Logic Theorist, primul model computaţional al cogniţiei umane şi, în acelaşi timp, primul sistem de inteligenţă artificială, a fost dezvoltat de Herbert Simon, de la Carnegie Tech, şi de Allen Newell, de la RÂND Corporation. Logic Theorist producea automat dovezi matematice emulând strategiile utilizate de matematicienii umani.

Provocarea în dezvoltarea modelelor computaţionale ale cogniţiei umane este să faci modele care sunt atât precise, cât şi generalizabile. Un model precis, bineînţeles, prognozează comportamentul uman cu un minimum de erori. Un model generalizabil poate face predicţii într-o gamă variată de circumstanţe, inclusiv circumstanţe neanticipate de creatorii săi – de exemplu, un model bun pentru clima Pământului ar trebui să poată prognoza consecinţele creşterii temperaturii globale, chiar dacă acesta nu este un lucru pe care să-l fi luat în considerare oamenii de ştiinţă care l-au proiectat. Însă, când vine vorba despre înţelegerea minţii omeneşti, aceste două obiective – acurateţea şi capacitatea de a generaliza – au fost tot timpul în dezacord.

La extrema capacităţii de a generaliza se află teoriile raţionale asupra cogniţiei. Aceste teorii descriu comportamentul uman ca răspuns raţional la o situaţie dată. Un actor raţional se străduieşte să maximizeze recompensa aşteptată produsă de o secvenţă de acţiuni – o idee cu utilizare largă în economie, *deoarece* produce astfel de prognoze generalizabile cu privire la comportamentul uman. Din acelaşi motiv, raţionalitatea este presupunerea standard a modelelor bazate pe învăţare cu întărire care încearcă să facă deducţii din comportamentul uman – poate sub rezerva că oamenii nu sunt agenţi perfect raţionali şi aleg uneori aleatoriu să acţioneze în feluri nealiniate sau chiar opuse celor mai bune interese ale lor.

Problema cu raţionalitatea ca bază de modelare a cogniţiei umane

este că nu este precisă. În domeniul luării deciziilor, un volum mare de literatură – condus de munca psihologilor cognitivi Daniel Kahneman şi Amos Tversky – a documentat modurile în care oamenii deviază de la prescripţiile modelelor raţionale. Kahneman şi Tversky au propus că, în multe situaţii, oamenii urmează în schimb o euristică simplă care le permite să ajungă la soluţii bune cu un cost cognitiv scăzut, dar uneori conduce la erori. Pentru a prelua unul dintre exemplele lor, dacă ceri cuiva să evalueze probabilitatea unui eveniment, se pot bizui pe cât de uşor este să generezi un exemplu pentru un astfel de eveniment din memorie, pot analiza dacă pot încropi o poveste cauzală pentru apariţia evenimentului sau pot evalua cât este de similar evenimentul cu aşteptările lor. Fiecare euristică este o strategie rezonabilă de a evita calcule probabilistice complexe, dar în acelaşi timp conduce şi la erori. De exemplu, dacă te bazezi pe cât de uşor este să generezi un eveniment din memorie ca ghid pentru probabilitatea sa, subestimezi şansele unor evenimente extreme (care sunt prin natura lor extrem de memorabile), cum ar fi atacurile teroriste.

Euristica pune la dispoziţie un model mai precis al cogniţiei umane, dar unul care nu este uşor generalizabil. Cum ştim ce euristică ar putea folosi oamenii într-o anumită situaţie? Mai există şi alte euristici pe care le folosesc şi pe care noi nu le-am descoperit încă? A şti exact cum se vor comporta oamenii într-o situaţie nouă este o provocare: este această situaţie una în care generează exemple din memorie, inventează poveşti cauzale sau se bazează pe similitudini?

În cele din urmă, avem nevoie de o modalitate de a descrie cum funcţionează mintea umană care are generalizabilitatea raţionalităţii şi acurateţea euristicii. O cale de a atinge acest obiectiv este să începem cu raţionalitatea şi să ne gândim cum o putem duce într-o direcţie mai realistă. O problemă cu utilizarea raţionalităţii ca bază pentru descrierea comportamentului unui agent real este că, în multe situaţii, calcularea acţiunilor raţionale necesită ca agentul să posede o cantitate uriaşă de resurse computaţionale. Poate să renteze să cheltuieşti acele resurse dacă iei o decizie cu consecinţe majore şi ai mult timp pentru a-ţi evalua opţiunile, dar cele mai multe decizii omeneşti sunt luate rapid şi pentru mize relativ mici. În orice caz, când timpul consumat pentru luarea unei decizii este costisitor – dacă nu din alte motive, măcar pentru că l-ai putea petrece făcând altceva –, noţiunea clasică de raţionalitate nu mai este o reţetă bună pentru comportamentul oamenilor.

Pentru a dezvolta un model mai realist al comportamentului raţional, trebuie să luăm în considerare costul computaţional. Agenţii reali trebuie să-şi moduleze timpul pe care-l petrec gândindu-se prin efectul pe care gândirea suplimentară îl are asupra rezultatelor unei

decizii. Dacă încerci să alegi o periuţă de dinţi, probabil nu trebuie să analizezi toate cele patru mii de rezultate care apar pe Amazon.com pentru periuţe manuale înainte de a cumpăra una: schimbi timpul pe care-l petreci căutând pe diferenţa pe care o face în calitatea rezultatului. Acest schimb poate fi formalizat, rezultând într-un model de comportament raţional pe care cercetătorii inteligenţei artificiale l-au numit „opdmalitate limitată”[49]. Agentul optim limitat nu se concentrează pe a alege întotdeauna acţiunea perfectă, ci mai degrabă pe găsirea algoritmului perfect de urmat pentru a putea găsi echilibrul perfect între a face greşeli şi a gândi prea mult.

Optimalitatea limitată face legătura între raţionalitate şi euristică. Descriind comportamente ca rezultat al unei alegeri raţionale cu privire la cât de mult timp să acordăm gândirii, oferă o teorie generalizabilă – adică una care poate fi aplicată în situaţii noi. Uneori, strategiile simple care au fost identificate ca euristice şi pe care le urmează oamenii se dovedesc a fi soluţii optime limitate. Prin urmare, decât să condamnăm euristicile folosite de oameni ca iraţionale, le putem privi ca pe un răspuns raţional la restricţiile impuse asupra calculului.

Dezvoltarea optimalităţii limitate ca teorie a comportamentului uman este un proiect în desfăşurare pe care echipa mea de cercetare şi alţii îl urmăresc activ. Dacă aceste eforturi dau roade, ne vor pune la dispoziţie cel mai important ingredient de care avem nevoie pentru a face sistemele de inteligenţă artificială mai inteligente atunci când încearcă să interpreteze acţiunile oamenilor, permiţând un model generativ pentru comportamentul oamenilor.

Luarea în considerare a restricţiilor computaţionale care intră în ecuaţia cogniţiei umane va fi foarte importantă atunci când vom începe să dezvoltăm sisteme automate care nu mai sunt supuse aceloraşi restricţii. Imaginează-ţi un sistem superinteligent de inteligenţă artificială care încearcă să-şi dea seama de ce le pasă oamenilor. Vindecarea cancerului sau confirmarea ipotezei Riemann, de exemplu, nu vor părea pentru o astfel de inteligenţă artificială lucruri care să fie foarte importante pentru noi: dacă aceste soluţii sunt evidente pentru un sistem superinteligent, şi-ar putea pune întrebarea de ce nu le-am găsit noi înşine şi ar putea trage concluzia că aceste probleme nu înseamnă mare lucru pentru noi. Dacă ne-ar fi păsat şi problemele ar fi fost atât de simple, le-am fi rezolvat deja. O deducţie rezonabilă ar fi că ne preocupăm cu ştiinţa şi cu matematica doar pentru că ne plac, şi nu pentru că ne pasă de rezultate.

Oricine are copii poate aprecia problema de a încerca să interpretezi comportamentul unui agent supus unor restricţii computaţionale diferite faţă de ale noastre. Părinţii bebeluşilor pot pierde ore întregi încercând să descurce adevăratele motivaţii care

stau la baza unor comportamente aparent inexplicabile. În calitate de tată şi cercetător în domeniul cogniţiei, am constatat că este mai uşor să înţeleg crizele bruşte ale fetiţei mele de doi ani când accept că se află la vârsta în care conştientizează că oamenii au dorinţe diferite, dar nu conştientizează că alţi oameni nu ştiu care sunt dorinţele ei. Este uşor de înţeles, apoi, de ce se enerva când alţi oameni nu făceau ce-şi dorea ea (aparent în mod transparent). Pentru a înţelege bebeluşii, trebuie să construieşti un model cognitiv al minţii bebeluşului. Sistemele superinteligente de inteligenţă artificială fac faţă aceleiaşi provocări când încearcă să înţeleagă comportamentul omenesc.

Probabil va mai trece mult timp până vom avea o inteligenţă artificială superinteligentă. Pe termen scurt, dezvoltarea unor modele mai bune pentru oameni se poate dovedi extrem de valoroasă pentru orice companie care câştigă bani analizând comportamentul omenesc – adică, în prezent, aproape orice companie care face afaceri pe web. În ultimii câţiva ani, tehnologii comerciale importante pentru interpretarea imaginilor şi textului au fost create din dezvoltarea unor modele mai bune pentru vedere şi limbaj. Dezvoltarea unor modele bune pentru oameni este următoarea bornă importantă.

Bineînţeles, dacă înţelegem cum funcţionează mintea omenească, îmbunătăţirea interacţiunii oamenilor cu computerele nu este singurul avantaj. Compromisul dintre a face greşeli şi a gândi prea mult ce caracterizează cogniţia omenească este un compromis pe care trebuie să-l facă toţi agenţii inteligenţi din lumea reală. Oamenii sunt un exemplu uimitor de sisteme care acţionează inteligent în ciuda unor restricţii computaţionale semnificative. Suntem foarte buni când trebuie să dezvoltăm strategii care ne permit să rezolvăm probleme fără a munci prea mult. Când vom înţelege cum facem acest lucru, vom reuşi să facem computerele să lucreze mai inteligent, nu mai mult.

Capitolul 13

Integrarea omului în ecuaţia inteligenţei artificiale

Anca Drăgan

Anca Drăgan este profesor asistent la Departamentul de Inginerie Electrică şi Informatică de la UC Berkeley.

A coanfiinţat şi se află în comitetul de conducere al Laboratorului Berkeley Al Research (BAIR) şi este coinvestigator principal la Centrul de Inteligenţă Artificială Compatibilă cu Omul, de la Universitatea Berkeley.

Cercetarea româncei Anca Drăgan se concentrează pe algoritmi care vor permite roboţilor să lucreze cu, în jurul şi în sprijinul oamenilor. Conduce Laboratorul Inter ACT de la Berkeley, unde studenţii săi lucrează la mai multe aplicaţii, de la roboţi ce oferă asistenţă la fabricaţie până la maşini autonome şi îşi bazează activitatea pe controlul optim, planificare, estimare, învăţare şi ştiinţe cognitive. Având treizeci şi ceva de ani, a scris deja în colaborare câteva lucrări cu colegul veteran de la Berkeley şi mentorul său Stuart Russell, abordând diverse aspecte ale învăţării automate şi problemele noduroase ale alinierii valorilor.

Împărtăşeşte preocuparea lui Stuart privind siguranţa inteligenţei artificiale. „Un risc imediat sunt agenţii care produc comportamente nedorite, surprinzătoare", a spus ea unui intervievator de la Institutul Future of Life. „Chiar dacă plănuim să folosim inteligenţa artificială pentru bine, lucrurile pot lua o turnură greşită, în special pentru că nu ne pricepem să specificăm obiectivele şi restricţiile agenţilor de inteligenţă artificială. Soluţiile lor sunt adesea ceva ce nu ne convine".

Obiectivul său principal este, aşadar, să ajute roboţii şi pe programatori să depăşească multele conflicte ce iau naştere din cauza lipsei de transparenţă cu privire la intenţiile celuilalt. Roboţii, spune ea, trebuie să ne pună întrebări. Trebuie să-şi pună singuri întrebări în legătură cu sarcinile primite şi trebuie să stea pe capul programatorilor umani până când toată lumea este de acord – pentru a evita ceea ce a numit ea eufemistic „efecte secundare nedorite".

La baza inteligenţei artificiale se află definiţia noastră matematică pentru ce înseamnă un agent al inteligenţei artificiale (un robot). Când definim un robot, definim stări, acţiuni şi recompense. Să ne gândim la un robot care face livrări, de exemplu. Stările sunt locuri din lume, iar acţiunile sunt mişcările pe care robotul le face pentru a ajunge dintr-o poziţie în alta apropiată. Pentru a permite robotului să decidă ce acţiune urmează să facă, definim o funcţie de recompensare – o mapare a stărilor şi acţiunilor în punctaje care arată cât de bună a fost acţiunea respectivă în cadrul stării – şi facem robotul să aleagă acţiunile care ar acumula maximul de „recompensă". Robotul obţine o recompensă mare când ajunge la destinaţie şi implică un mic cost de fiecare dată când se mişcă; această funcţie de recompensare stimulează robotul să ajungă la destinaţie cât se poate de repede. Similar, o maşină autonomă ar putea obţine o recompensă făcând progrese pe ruta sa şi ar putea implica un cost dacă se apropie prea mult de alte maşini.

Date fiind aceste definiţii, sarcina unui robot este să-şi dea seama ce acţiuni trebuie să întreprindă pentru a obţine cea mai mare

recompensă cumulată. Am lucrat din greu în domeniul inteligenţei artificiale pentru a permite roboţilor să facă acest lucru. Implicit, am presupus că, dacă vom reuşi – dacă roboţii ar putea lua definiţia unei probleme şi ar putea-o transforma într-o lege în baza căreia să acţioneze –, vom obţine roboţi care sunt utili oamenilor şi societăţii.

Până acum, nu ne-am înşelat prea mult. Dacă vrei o inteligenţă artificială care să clasifice celulele drept canceroase sau benigne sau un robot care aspiră covorul din sufragerie când eşti la serviciu, avem ce să-ţi oferim. Unele probleme reale pot fi definite într-adevăr în izolare, cu stări, acţiuni şi recompense clare. Dar cu capacităţile crescânde ale inteligenţei artificiale, problemele pe care vrem să le abordăm nu intră atât de uşor în acest cadru. Nu mai putem tăia o bucăţică minusculă din lume, după care să o aşezăm frumos într-o cutie şi să i-o oferim unui robot. A-i ajuta pe oameni începe să însemne a lucra în lumea reală, unde trebuie să interacţioneze fizic cu oamenii şi să tragă concluzii legate de aceştia. „Oamenii" vor trebui să intre formal undeva în definiţia problemei inteligenţei artificiale.

Maşinile autonome au fost dezvoltate deja. Vor trebui să împartă drumul cu vehicule care sunt conduse de oameni şi cu pietoni şi vor trebui să înveţe să facă astfel compromisuri între a ne aduce acasă cât se poate de repede şi a respecta nevoile celorlalţi şoferi. Asistenţii personali vor trebui să-şi dea seama când avem nevoie de ajutor şi în ce măsură şi ce sarcini preferăm să ducem la bun sfârşit noi înşine versus ce am prefera să facă ei. Un sistem de asistenţă pentru decizii sau un sistem de diagnosticare medicală va trebui să ne explice recomandările pentru a le putea înţelege şi verifica. Meditatorii automaţi vor trebui să stabilească ce exemple sunt informative sau ilustrative – nu pentru alte maşini, ci pentru noi, oamenii.

Privind şi mai departe în viitor, dacă vrem ca inteligenţele artificiale foarte capabile să fie compatibile cu oamenii, nu le putem crea izolate de oameni, după care să încercăm să le facem compatibile; mai degrabă, va trebui să definim inteligenţa artificială „compatibilă cu omul" încă de la început. Oamenii nu pot fi o idee care vine mai târziu.

Când vine vorba despre roboţi reali care ajută oameni reali, definiţia standard a inteligenţei artificiale nu ne ajută din două motive fundamentale: primul, optimizarea funcţiei de recompensă a robotului în izolare este diferită de optimizarea ei când robotul acţionează pe lângă oameni, pentru că şi oamenii acţionează. Luăm decizii în baza propriilor noastre interese, şi aceste decizii dictează ce acţiuni putem executa. În plus, raţionăm cu privire la robot, adică răspundem la ceea ce credem că face sau va face şi la capacităţile pe care credem noi că le are. Acţiunile pe care le face robotul trebuie să se integreze bine cu ale noastre. Aceasta este *problema coordonării*.

În al doilea rând, omul decide în cele din urmă care va fi funcţia de recompensă a robotului. Şi recompensa are ca scop stimularea comportamentului robotului care corespunde cu ce doreşte clientul final, designerul sau chiar societatea întreagă. Consider că roboţii capabili, care trec dincolo de sarcinile restrânse şi definite clar, vor trebui să înţeleagă acest lucru pentru a obţine compatibilitatea cu oamenii. Aceasta este problema *alinierii valorilor*.

Problema coordonării: oamenii nu sunt simple obiecte în mediul înconjurător

Când proiectăm roboţi pentru o anumită sarcină, este tentant să-i abstractizăm pe oameni până la uitare. Un asistent personal robotic, de exemplu, trebuie să ştie cum să se mişte pentru a ridica obiecte, aşa că definim această problemă în izolare faţă de oamenii pentru care robotul urmează să ridice obiecte. Totuşi, în timp ce robotul se mişcă, nu vrem să se izbească de lucruri sau de oameni, aşa că am putea include localizarea fizică a persoanei în definiţia stării robotului. La fel şi pentru maşini: nu vrem să se ciocnească cu alte maşini, aşa că le permitem să urmărească poziţiile celorlalte maşini şi să presupună că se vor deplasa uniform, în aceeaşi direcţie, pe viitor. O fiinţă umană, în acest sens, nu este diferită pentru un robot faţă de o minge care se rostogoleşte pe o suprafaţă plată. Mingea se va comporta în următoarele câteva secunde în acelaşi fel în care s-a comportat şi în ultimele secunde, continuă să se rostogolească în aceeaşi direcţie cu aproximativ aceeaşi viteză. Bineînţeles, aceasta nu seamănă cu comportamentul oamenilor, dar o astfel de simplificare permite multor roboţi să-şi ducă sarcinile la bun sfârşit şi, în cea mai mare parte, să nu-i încurce pe oameni. Un robot din casa ta, de exemplu, te poate vedea venind pe hol, se poate feri din calea ta pentru a te lăsa să treci şi îşi poate relua sarcina de unde a rămas după ce treci.

Pe măsură ce roboţii au devenit din ce în ce mai capabili însă, tratarea oamenilor ca obiecte în continuă mişcare începe să nu mai fie eficientă. Un şofer uman care schimbă banda nu va continua în aceeaşi direcţie, dar va continua să meargă în faţă după ce a schimbat banda. Când încerci să ajungi la un obiect, adesea te întinzi după el printre alte obiecte şi te opreşti când ajungi la cel pe care-l doreşti. Când mergi pe un hol, ai o destinaţie în minte: poţi face dreapta şi intra în dormitor sau poţi face stânga şi intra în sufragerie. Presupunerea că nu suntem diferiţi de o minge care se rostogoleşte conduce la ineficienţă atunci când robotul se fereşte dacă nu este necesar şi poate pune robotul în pericol când comportamentul persoanei se schimbă. Doar şi pentru a nu-i încurca pe oameni, roboţii trebuie să fie oarecum precişi în anticiparea acţiunilor umane. Şi, spre deosebire de mingea care se rostogoleşte, ce fac oamenii depinde de ceea ce *decid* să facă. Prin urmare, pentru a anticipa acţiunile omeneşti, roboţii trebuie să înceapă să înţeleagă procesul de luare a deciziilor. Şi asta nu înseamnă să presupună despre comportamentul uman că este în întregime optim; poate ar fi suficient pentru un robot care joacă şah sau Go, dar, în lumea reală, deciziile oamenilor sunt mai puţin predictibile decât mutarea optimă pe o tablă de joc.

Această nevoie de a înţelege acţiunile şi deciziile umane se aplică

atât roboţilor fizici, cât şi celor nefizici. Dacă oricare dintre cele două tipuri îşi bazează deciziile despre cum să acţioneze în legătură cu presupunerea că un om poate face un lucru şi face altul, neconcordanţa rezultată poate fi catastrofală. Pentru maşini, poate însemna coliziune. Pentru o inteligenţă artificială cu un rol economic, de exemplu, neconcordanţa între ce se aşteaptă să facem şi ce facem efectiv poate avea consecinţe şi mai rele.

O alternativă este ca robotul să nu prezică acţiunile umane şi să se protejeze, în schimb, de *cel mai rău caz posibil* creat de o acţiune omenească. Adesea, când roboţii fac asta însă, nu mai sunt utili. Cu maşinile, acest rezultat este blocat, pentru că face totul prea riscant.

Toate acestea ne pun pe noi, comunitatea cercetătorilor inteligenţei artificiale, într-o situaţie delicată. Sugerează că roboţii vor avea nevoie de modele predictive precise (sau, cel puţin, rezonabile) pentru orice vor decide oamenii să facă. Modul în care încercăm să definim stările nu poate include poziţia fizică a oamenilor în lume. În schimb, trebuie să estimăm ceva *intern* pentru oameni. Va trebui să proiectăm roboţi ce ţin cont de această stare internă a oamenilor, şi asta nu va fi uşor. Din fericire, oamenii tind să ofere indicii roboţilor cu privire la starea lor internă: acţiunile lor în desfăşurare oferă robotului observaţii (în sensul deducţiilor beyesiene) cu privire la intenţiile lor. Dacă începem să mergem pe partea dreaptă a holului, probabil vom intra în următoarea încăpere de pe partea dreaptă.

Ce face această problemă şi mai complicată este faptul că oamenii nu iau decizii în izolare. Ar fi ceva dacă roboţii ar putea preconiza acţiunile pe care urmează să le facă o persoană şi să-şi dea seama doar ce au de făcut ca răspuns. Dar, din păcate, aceasta ar conduce la roboţi ultradefensivi care ar bulversa complet oamenii. (Gândeşte-te la şoferii umani care se blochează în intersecţii cu patru sensuri de mers, de exemplu.) Ce-i lipseşte abordării bazate pe preconizarea intenţiei este că, în momentul în care robotul acţionează, acea acţiune a sa influenţează acţiunile omului.

Există o influenţă reciprocă între roboţi şi oameni, una prin care roboţii vor trebui să înveţe să navigheze. Problema nu este întotdeauna planificarea robotului în jurul oamenilor; şi oamenii planifică în jurul robotului. Este important ca roboţii să ţină cont de acest lucru când decid acţiunile lor ulterioare, indiferent dacă se află pe drum, în bucătărie sau chiar în spaţii virtuale, unde o acţiune poate însemna o achiziţie sau adoptarea unei noi strategii. Acest lucru ar conferi roboţilor strategii *de coordonare*, permiţându-le să ia parte la negocierile pe care oamenii le fac fără efort în fiecare zi – de la persoana care are prioritate la intersecţie sau printr-o uşă îngustă la rolul pe care ni-l asumăm când colaborăm la pregătirea micului dejun până la obţinerea unui consens cu privire la pasul următor dintr-un

proiect.

În cele din urmă, la fel cum roboţii trebuie să anticipeze ce vor face oamenii, şi oamenii trebuie să procedeze la fel cu roboţii. Acesta este motivul pentru care transparenţa este importantă. Roboţii nu vor avea nevoie numai de modele mentale bune pentru oameni, dar şi oamenii vor avea nevoie de modele mentale bune pentru roboţi. Modelul pe care-l are o persoană pentru un robot trebuie să fie integrat în definiţia noastră pentru stare, iar robotul trebuie să fie conştient de modul în care acţiunile sale modifică modelul respectiv. Asemănător robotului care tratează acţiunile umane ca indicii pentru stările interne ale omului, oamenii îşi vor schimba părerile despre robot pe măsură ce-i observă acţiunile. Din păcate, roboţii nu dau indicii la fel de uşor ca oamenii; noi am exersat mult să comunicăm implicit cu oamenii. Dar a permite roboţilor să conştientizeze modificarea pe care o cauzează acţiunile lor în modelul mental al persoanei pentru robot poate conduce la acţiuni mai atent alese care conferă indicii corecte – care comunică în mod clar oamenilor intenţiile robotului, funcţia sa de recompensă şi limitările sale. De exemplu, un robot îşi poate schimba mişcarea când cară ceva greu pentru a sugera dificultatea pe care o are în manevrarea obiectelor grele. Cu cât oamenii ştiu mai multe despre robot, cu atât este mai uşor să se coordoneze cu el.

Obţinerea compatibilităţii în acţiuni va necesita ca roboţii să anticipeze acţiunile omeneşti, să ţină cont de modul în care acele acţiuni le vor influenţa pe ale lor şi să permită oamenilor să anticipeze acţiunile roboţilor. Cercetarea a făcut progrese în vederea depăşirii acestor obstacole, dar mai avem un drum lung de parcurs.

Problema alinierii valorii: oamenii deţin cheia funcţiei de recompensă a roboţilor

Progresul care va permite roboţilor să optimizeze recompensa pune o povară mai grea pe umerii noştri, ai designerilor, pentru a le conferi recompensa potrivită care trebuie optimizată în primul rând. Ideea iniţială era că, pentru orice sarcină pe care voiam să i-o dăm robotului, am putea scrie o funcţie de recompensare ce oferă stimulente pentru comportamentul potrivit. Din păcate, ce se întâmplă adesea este că specificăm o anumită funcţie de recompensare şi comportamentul care reiese din optimizarea ei nu este ceea ce vrem. Funcţiile de recompensare intuitivă, atunci când se combină cu instanţe neobişnuite ale unei sarcini, pot conduce la comportamente neintuitive. Recompensezi un agent dintr-un joc de curse cu un punctaj în joc şi, în unele cazuri, găseşte o portiţă pe care o exploatează pentru a câştiga mult mai multe puncte fără a câştiga efectiv cursa. Stuart Russell şi Peter Norvig oferă un exemplu minunat

în cartea lor *Artificial Intelligence*: *A Modern Approach*: recompensarea unui robot de aspirat pentru cât de mult praf aspiră va face ca acesta să scoată din praful deja aspirat şi să-l aspire din nou pentru a obţine o recompensă mai mare.

În general, oamenilor le-a fost foarte greu să specifice exact ce vor, după cum arată toate acele poveşti cu djinni. O paradigmă a inteligenţei artificiale în care roboţii pot obţine recompense specificate extern eşuează atunci când recompensa nu este gândită perfect. Poate oferi stimulente robotului să se comporte greşit şi chiar să reziste încercărilor noastre de a-i corecta comportamentul, pentru că această corecţie ar conduce la o recompensă mai mică.

O paradigmă aparent mai bună ar fi ca roboţii să optimizeze ceea ce ne dorim intern, chiar dacă ne este mai greu să explicăm. Aceştia ar urma să folosească ce spunem şi ce facem ca dovezi pentru ceea ce vrem mai degrabă decât a interpreta asta literal sau ca pe un dat. Când scriem o funcţie de recompensare, robotul ar trebui să înţeleagă că ne putem înşela; că este posibil să nu fi luat în considerare toate faţetele sarcinii; că nu există nicio garanţie că funcţia de recompensare dată va conduce *întotdeauna* la comportamentul pe care ni-l dorim. Robotul ar trebui să integreze ce am scris noi în înţelegerea sa cu privire la ce ne dorim noi, dar ar trebui şi să distileze informaţii clarificatoare din comportamentul nostru. Ar trebui să urmărească ghidarea noastră, pentru că acesta este singurul mod de a optimiza funcţia de recompensare dorită cu adevărat.

Chiar dacă reuşim să conferim roboţilor capacitatea de a învăţa ce dorim, rămâne o întrebare importantă la care inteligenţa artificială nu poate răspunde singură. Putem face roboţii să încerce să-şi alinieze valorile cu valorile interne ale unei persoane, dar problema este că e vorba despre mai multe persoane. Robotul are un utilizator final (sau, poate, câţiva, cum ar fi un robot personal care îngrijeşte o familie, o maşină care conduce mai mulţi pasageri în mai multe destinaţii sau un asistent de birou pentru o echipă întreagă); are un designer (sau, poate, câţiva); şi interacţionează cu societatea – maşina autonomă împarte drumul cu pietoni, vehicule conduse de oameni şi alte maşini autonome. Cum vom combina valorile acestor entităţi când ar putea intra în conflict este o problemă importantă pe care trebuie să o rezolvăm. Cercetarea din domeniul inteligenţei artificiale ne poate oferi instrumentele necesare pentru a combina valorile în orice fel decidem, dar nu poate lua deciziile necesare pentru noi.

Pe scurt, trebuie să permitem roboţilor să emită judecăţi în privinţa noastră – să ne vadă ca pe ceva mai mult decât simple obstacole sau jucători perfecţi. Trebuie să ne ia natura umană în considerare, astfel încât să fie bine coordonaţi şi bine aliniaţi cu noi. Dacă reuşim, vom avea într-adevăr instrumente care ne vor spori

substanţial calitatea vieţii.

Capitolul 14
Coborârea pe panta unui gradient

Chris Anderson

Chris Anderson este antreprenor; fost redactor-şef al revistei Wired, cofondator şi CEO al companiei 3DR şi autorul cărţilor The Long Tailf Free *şi* Makers.

Compania lui Chris Anderson, 3DR, a ajutat la lansarea industriei moderne a dronelor şi se concentrează acum pe softul datelor pentru drone. S-a lansat construind o comunitate de robotică aeriană cu sursă deschisă denumită DIY Drones[50] şi a făcut câteva experimente timpurii nerecomandate, cum ar fi să bâzâie la Laboratorul Lawrence Berkeley cu unul dintre spionii săi zburători autonomi. Putea foarte bine să fie un caz de expresie a genelor antice, de vreme ce este descendentul unui fondator al mişcării anarhiste din America. Chris a condus revista Wired, publicaţia preferată a tehno-utopiştilor, dar şi a tehno-distopiştilor, din 2001 până în 2012; în timpul mandatului său, revista a câştigat cinci premii National Magazine.

Chris nu agreează termenul „robotician" („ca orice robotician modest, eu nu mă consider unuL). A început ca fizician. „M-am dovedit a fi un fizician slaba, mi-a spus el recent. „M-am luptat, am mers la Los Alamos şi m-am gândit «Ei bine, poate că n-o să fiu laureat al Premiului Nobel, dar pot să fiu totuşi un om de ştiinţă». Toţi cei care eram în domeniul fizicii şi aveam aceşti eroi romantici – fraţii Feynman, proiectul Manhattan – ne-am dat seama că traiectoria carierei noastre ne-ar duce în cel mai bun caz la un proiect de la CERN în cincisprezece ani. Acel proiect urma să fie un eşec, caz în care nu ar apărea nicio lucrare, sau un succes, caz în care ai fi autorul numărul trei sute pe lucrarea respectivă şi ai deveni profesor asistent la Universitatea de Stat din Iowa".

„Majoritatea colegilor mei au mers pe Wall Street şi au devenit quanţi[51], şi acestora le datorăm criza imobiliarelor. Alţii au pus bazele internetului. Mai întâi, am construit internetul conectând laboratoarele de fizică. Apoi am construit webul. Apoi am fost primii care am folosit Big Data. Aveam supercomputere – Crays – care aveau jumătate din puterea telefoanelor noastre de acum, dar erau supercomputerele vremii. Între timp, citeam această revistă numită Wired, care a apărut pe piaţă în 1993, şi ne-am dat seama că acest instrument pe care-l folosim noi, oamenii de ştiinţă, ar

putea fi folosit de toată lumea. Internetul nu fusese creat doar pentru date ştiinţifice, era o revoluţie culturală uluitoare. Aşa că, atunci când Condé

Nast mi-a cerut să preiau conducerea revistei, am spus «Clar!». Această revistă mi-a schimbat viaţa".

Avea cinci copii la vremea aceea – toţi jucau jocuri video – care l-au făcut să fie interesat de „roboţii zburători". Şi-a dat demisia de la Wired. Restul este o istorie din Silicon Valley.

Viaţa

Ţânţarul îmi detectează mirosul de la zece metri. Mirosul meu îi declanşează funcţia de urmărire, care constă în cele mai simple reguli posibile. Mai întâi, deplasarea într-o direcţie aleatorie. Dacă mirosul se intensifică, continuă să se deplaseze în direcţia respectivă. Dacă mirosul scade, se deplasează în partea opusă. Dacă mirosul dispare, se mişcă în lateral până când îl regăseşte. Repetă acest set până la atingerea ţintei.

Norul mirosului meu este cel mai dens lângă mine şi dispare pe măsură ce distanţa creşte, o ceaţă invizibilă de particule emanată din pielea mea care se mişcă precum fumul în vânt. Cu cât este mai aproape de pielea mea, cu atât este mai mare densitatea particulelor; cu cât este mai departe, cu atât este mai mică. Scăderea este numită gradient, cuvânt ce descrie o tranziţie progresivă de la un nivel la altul – în opoziţie cu funcţia continuă, care descrie o schimbare discretă.

Când ţânţarul urmăreşte acest gradient până la sursă folosindu-şi algoritmul simplu, aterizează pe pielea mea, pe care o simte cu senzorii termici din picioare, care sunt acordaţi la alt gradient: temperatura. Apoi îşi împinge trompa sub formă de ac prin suprafaţă, unde al treilea set de senzori din vârful acesteia detectează alt gradient, acela al densităţii sangvine. Acul flexibil se mişcă pe sub pielea mea până când mirosul de sânge îl ghidează către un vas capilar, pe care îl perforează. Apoi sângele meu începe să curgă în ţânţar. Misiune îndeplinită. Au!

Ce pare un radar puternic al insectelor în întuneric, cu o inteligenţă inexplicabilă pentru nişte creiere atât de mici în ceea ce priveşte căutarea sângelui, este, de fapt, doar un nas sensibil, lipsit aproape complet de inteligenţă. Ţânţarii sunt mai apropiaţi de plantele care caută soarele decât de rachetele ghidate. Însă, dacă aplicăm literal această regulă „urmează-ţi nasul", se pot deplasa prin toată casa şi te pot găsi, se pot strecura prin crăpăturile din uşa cu plasă şi pot chiar identifica bucăţica de piele pe care ai lăsat-o neacoperită dintre pălărie şi gulerul cămăşii. Este doar un mers aleatoriu, combinat cu aripi flexibile şi picioare care permit insectelor să ricoşeze din obstacole, precum şi un instinct de a coborî pe panta

unui gradient chimic.

Dar „coborârea pe panta unui gradient” este mult mai mult decât metoda de navigare a insectelor. Priveşte în jurul tău şi o vei vedea peste tot, de la cele mai simple reguli fizice ale universului până la cea mai avansată inteligenţă artificială.

Universul

Trăim într-o lume compusă din nenumăraţi gradienţi, de la lumină şi căldură la gravitaţie şi dâre chimice (dâre chimice![52]) Apa curge de-a lungul unui gradient al gravitaţiei, iar corpul tău trăieşte din soluţii chimice care alunecă pe membranele celulare de la o concentraţie mare înspre una mică. Toate acţiunile din univers au la bază un motor gradient, de la mişcarea planetelor de-a lungul gradienţilor gravitaţionali până la fuziunea atomilor de-a lungul gradienţilor cu încărcătură electrică pentru a forma molecule. Propriile noastre trebuinţe, cum ar fi foamea şi somnul, au ca motor gradienţi electrochimici aflaţi în corpurile noastre. Şi funcţiile creierului nostru, semnale electrice care se deplasează pe canalele ionizate din sinapsele ce ne conectează neuronii, simt pur şi simplu atomi şi electroni care curg „în aval” pe alţi gradienţi electrici şi chimici. Uită de analogiile cu ceasurile, creierele noastre sunt mai aproape de un sistem de canale şi încuietori, cu semnale care se deplasează ca apa dintr-o stare în alta.

În timp ce stau aici şi scriu aceste rânduri, caut, de fapt, stări de echilibru într-o topologie «-dimensională de gradienţi. Să ne gândim doar la unul: căldura. Temperatura corpului meu este mai ridicată decât temperatura ambientală, aşa că radiez căldură, care trebuie înlocuită în corpul meu. Chiar şi bacteriile din tractul meu digestiv folosesc senzori care măsoară concentraţiile de zahăr din lichidul ce le înconjoară şi dau din cilii lor asemănători unei cozi pentru a înota „în amonte”, unde depozitele de zahăr sunt cele mai bogate. Starea naturală a tuturor sistemelor este să curgă către stările cu energie mai scăzută, un proces care este descris în sens larg ca entropie (tendinţa lucrurilor de a aluneca din stări de organizare în stări de dezorganizare; totul se va destrăma în cele din urmă, inclusiv Universul însuşi).

Dar cum explicăm comportamentele complexe, cum ar fi capacitatea noastră de a lua decizii? Răspunsul este în continuare coborârea pe panta unui gradient.

Creierele noastre

Pe cât este de miraculoasă şi enigmatică inteligenţa omenească, ştiinţa începe să accepte ideea despre creierele noastre că ar funcţiona

în acelaşi fel cu alte sisteme complexe cu straturi şi cicluri de *feedbacky* toate urmărind ceea ce numim matematic „funcţii de optimizare”, dar am putea la fel de bine să le numim „curgere în aval” într-un anumit sens.

Esenţa inteligenţei este învăţarea şi învăţăm corelând informaţiile de intrare cu punctaje pozitive sau negative (recompense sau pedepse). Prin urmare, pentru un bebeluş, „acest sunet” (vocea mamei tale) este asociat cu alte conexiuni învăţate cu „mama ta”, cum ar fi mâncarea şi confortul. Asemenea „această mişcare a muşchiului îmi aduce degetul mai aproape de gură”. În timp, prin încercare şi eroare, reţeaua neuronală a creierului întăreşte aceste conexiuni. Între timp, „această mişcare a muşchiului nu îmi aduce degetul mai aproape de gură” este o corelaţie negativă, iar creierul va slăbi acele conexiuni.

Însă acest model este prea simplist. Limitele coborârii pe panta unui gradient constituie aşa-numita problemă „minima locală” (sau problema „maxima locală” dacă avem de-a face cu *urcarea* pe panta unui gradient). Dacă mergi printr-o regiune muntoasă şi vrei să ajungi acasă şi dacă mergi tot timpul în jos, vei ajunge cu siguranţă în următoarea vale, dar nu neapărat şi peste ceilalţi munţi care sunt în jurul ei şi care se află între tine şi casă. Pentru a ajunge la destinaţie, ai nevoie de un model mental (de exemplu, o hartă) al topologiei, pentru a şti unde trebuie să urci şi să ieşi de pe vale, sau trebuie să alternezi între coborâri pe panta gradientului şi plimbări aleatorii pentru a putea ieşi din regiune.

Asta face, de fapt, ţânţarul când urmăreşte mirosul meu. Coboară când se află în norul meu şi se deplasează aleatoriu când pierde urma sau dă de un obstacol.

Inteligenţa artificială

Aşadar, asta-i natura. Dar computerele? Softurile tradiţionale nu funcţionează aşa – respectă arbori determinişti de logică dură: „dacă aşa, atunci fă aşa”. Dar softurile care interacţionează cu lumea fizică au tendinţa să funcţioneze *asemănător* lumii fizice. Asta înseamnă că au de-a face cu date de intrare zgomotoase (senzori sau comportament uman) şi oferă rezultate probabiliste, şi nu deterministe. Şi acest lucru, la rândul său, înseamnă şi mai mult o coborâre pe panta unui gradient.

Softurile inteligenţei artificiale sunt cele mai bune exemple, în special tipurile de inteligenţă artificială care utilizează modele artificiale de reţele neuronale (inclusiv reţele neuronale spiralate sau „profunde”, cu multe straturi). În acestea, un proces tipic constă în „instruirea” lor, arătându-le multe exemple pentru ce ne dorim să înveţe (imagini cu pisici care au eticheta „pisică” de exemplu), alături de exemple cu alte date aleatorii (imagini cu altceva). Această metodă

este denumită „învăţare supravegheată”, deoarece reţeaua neuronală este educată prin exemple, inclusiv utilizarea „instruirii adversative”, cu date care nu sunt corelate cu rezultatele dorite.

Aceste reţele neuronale, la fel ca modelele lor biologice, constau în straturi cu mii de noduri („neuroni”, în analogie), fiecare fiind conectat cu toate nodurile din straturile inferioare şi superioare prin conexiuni care au la început o forţă aleatorie. Nivelul superior primeşte date şi nivelul inferior generează răspunsul corect. Orice serie de conexiuni care s-a întâmplat să aterizeze pe răspunsul corect este întărită („recompensată”), iar cele care s-au înşelat sunt slăbite („pedepsite”). Dacă se repetă acest proces de zeci de mii de ori, avem de-a face cu o reţea complet instruită pentru acest tip de date.

Ne putem gândi la toate combinaţiile posibile de conexiuni ca la suprafaţa unei planete, cu dealuri şi văi. (Ignorăm momentan că suprafaţa este doar tridimensională şi topologia efectivă este multidimensională.) Optimizarea prin care trece reţeaua în timp ce învaţă este asemănătoare procesului de a găsi cea mai adâncă vale de pe planetă. Constă în următoarele etape:

1. *Definirea unei „funcţii de cost” care stabileşte cât de bine a rezolvat reţeaua problema.*
2. *Rularea reţelei o dată şi analiza rezultatului generat de funcţia de cost respectivă.*
3. *Modificarea valorilor conexiunilor şi repetarea procesului. Diferenţa dintre cele două rezultate este direcţia sau „panta” pe care reţeaua s-a deplasat între cele două încercări.*
4. *Dacă panta este „în jos”, conexiunile sunt modificate mai mult în acea direcţie. Dacă este „în sus”, sunt modificate în direcţia opusă.*
5. *Repetarea procedurii până când nu mai apare nicio îmbunătăţire în nicio direcţie. Ceea ce înseamnă că ne aflăm la punctul minim.*

Felicitări! Dar, probabil, este un minim *local* sau o mică adâncitură în munte, aşa că va trebui să continuăm dacă vrem ca rezultatul să fie şi mai bun. Nu mai putem continua în jos şi nu ştim unde este cel mai jos punct, aşa că trebuie să-l găsim cumva. Există mai multe posibilităţi, dar iată câteva:

1. *Încercăm de multe ori, cu setări aleatorii diferite, şi împărtăşim rezultatele fiecărei încercări; în fond, zgâlţâim sistemul pentru a vedea dacă se linişteşte undeva mai jos. Dacă unul dintre celelalte rezultate a găsit o vale mai adâncă, începem de la acele setări.*
2. *Nu mergem încontinuu în jos, ci şi „pe trei cărări”, ca un beţiv*

(metodă denumită „coborâre stocastică pe panta unui gradient"). Dacă facem asta suficient timp, vom găsi în cele din urmă punctul cel mai jos. E ca o metaforă pentru viaţă.

3. *Căutăm trăsături „interesante" definite de diversitate (limite sau schimbări de culoare, de exemplu). Avertizare: această metodă poate duce la nebunie – prea mult „interesant" atrage reţeaua către iluzii optice. Aşa că am face bine să rămânem sănătoşi şi să punem accent pe tipul de trăsături care prezintă o probabilitate mai mare de a fi reale în natură, spre deosebire de artefacte sau erori. Acest proces este numit „regularizare" şi există multe tehnici pentru a-l obţine, cum ar fi dacă acele caracteristici au mai fost văzute (învăţate) sau sunt „prea frecvente" (ca zgomotul de fundal), şi nu „mai puţin frecvente" (mai continue, ca trăsăturile reale din natură).*

Doar pentru că sistemele de inteligenţă artificială ajung uneori în valoarea minimă locală, să nu tragem concluzia că asta le face mai diferite de viaţă. Oamenii – şi, probabil, toate formele de viaţă – sunt adesea blocaţi în valoarea minimă locală.

Să luăm, de exemplu, înţelegerea noastră în ceea ce priveşte jocul de Go, care a fost predat, învăţat şi optimizat de oameni mii de ani. I-a luat inteligenţei artificiale trei ani să afle că l-am jucat greşit tot acest timp şi că există soluţii mai bune, aproape extraterestre, pentru acest joc pe care nu le-am luat niciodată în considerare – în cea mai mare parte, deoarece creierele noastre nu au puterea de procesare pentru a se gândi la atât de multe mutări în avans.

Chiar şi în şah, care este de zece ori mai simplu şi despre care credeam că l-am înţeles, maşinile cu forţă brută ne pot învinge cu propriile noastre strategii. Şi şahul, se pare, când este explorat de sisteme de inteligenţă artificială cu reţele neuronale superioare, are strategii bizare, dar superioare la care nu ne-am gândit, cum ar fi sacrificarea timpurie a reginei pentru a câştiga un avantaj obscur pe termen lung. Este ca şi cum am fi jucat versiuni bidimensionale ale unor jocuri care există în dimensiuni superioare.

Dacă ceva din ce spun îţi sună cunoscut, este pentru că fizica se luptă cu acest tip de probleme topologice de decenii întregi. Noţiunea de spaţiu fiind pluridimensională şi matematica redusă la înţelegerea geometriilor şi interacţiunilor „membranelor" care se află dincolo de simţurile noastre, este locul în care teoreticienii marii unificări merg să moară. Dar, spre deosebire de fizica teoretică multidimensională, inteligenţa artificială este ceva ce putem experimenta şi măsura.

Prin urmare, asta vom face. Următoarele câteva decenii vor fi o explorare explozivă a modalităţilor de gândire pe care 7 milioane de ani de evoluţie nu le-au găsit. Vom da din coate până ieşim din

valoarea minimă locală şi vom găsi valori minime mai profunde, poate chiar valoarea minimă globală. Şi, după ce le găsim, poate chiar învăţăm maşinile să pară la fel de inteligente ca un ţânţar, coborând încontinuu pe panta gradienţilor cosmici până la un ultim scop, oricare ar fi acesta.

Capitolul 15

„Informaţia" pentru Wiener, pentru Shannon şi pentru noi

David Kaiser

David Kaiser este profesor Germeshausen de istorie a ştiinţei şi profesor de fizică la MIT, director al programelor sale ştiinţifice, tehnologice şi sociale. Este autorul cărţilor How the Hippies Saved Physics: Science, Counterculture, and the Quantum Revival şi American Physics and the Cold Warbubble *(în curs de apariţie).*

David Kaiser este un fizician atipic, interesat de intersecţia ştiinţei sale cu politica şi cultura, despre care a scris mult.

La prima întâlnire (din Washington, Connecticut), care a precedat apariţia acestei cărţi, a comentat despre modul în care s-a modificat perspectiva asupra „informaţiei" de pe vremea lui Wiener: era complexului militar industrial din Războiul Rece. Pe atunci, Wiener compara informaţia, la nivel metaforic, cu entropia, în sensul că nu poate fi conservată, adică monopolizată; astfel, argumenta el, secretele noastre atomice şi alte probleme clasificate nu ar rămâne secrete prea mult timp. Astăzi, întrucât informaţia (aşa cum s-ar fi aşteptat Wiener), falsă sau nu, se scurge peste tot prin celălalt Washington, informaţiile din lumea economică au fost într-adevăr colectate, comunicate şi monetizate.

Această fixare, spune David, „nu era pe de-a-ntregul bună şi nici pe de-a-ntregul reaa – dacă, bănuiesc, te-ai săturat să fii sâcâit de reclame la şosete sau croaziere pe râuri europene în browserul tău la un minut după ce le-ai cumpărat deja.

Ca să nu spunem nimic despre proliferarea informaţiei. David ni s-a plâns nouă, participanţilor la întâlnire, că, pe vremea lui Wiener, fizicienii puteau „citi toată revista Physical Review. Stătea confortabil în faţa noastră într-o grămadă pe care o puteam gestiona. Acum suntem potopiţi cu cincizeci de mii de reviste cu sursă deschisă pe minut", pline cu Dumnezeu ştie ce. Wiener nu a anticipat niciuna dintre aceste dezvoltări, spune David, făcându-l să întrebe „Avem nevoie oare de un set nou de metafore după care să ne ghidăm?"

În *The Sleepwalkers*, o istorie cuprinzătoare a gândirii ştiinţifice din vremurile antice până în Renaştere, Arthur Koestler a identificat o tensiune care a marcat cele mai dramatice salturi ale imaginaţiei noastre cosmologice. Citind dintr-o perspectivă prezentă lucrările lui Nicolaus Copernicus şi Johannes Kepler, Koestler susţinea că suntem loviri de nefamiliaritatea lor bizară – integrarea lor în magia sau misticismul unei epoci trecute –, dar şi de observaţiile lor care par a fi moderne.

Eu văd aceleaşi sensuri duble – zigzagul de pliuri origami de vechi şi nou – în bine-cunoscuta carte a lui Norbert Wiener *The Human Use of Human Beings*. Publicată iniţial în 1950 şi revizuită în 1954, cartea este în multe sensuri extraordinar de profetică. Wiener, eruditul de la MIT, a recunoscut în faţa majorităţii observatorilor că „societatea poate fi înţeleasă doar prin studierea mesajelor şi instalaţiilor de comunicare ce îi aparţin". Wiener susţinea că ciclurile de *feedback*, elementul central din întreaga sa teorie cibernetică, vor urma să joace un rol central în dinamica socială. Aceste cicluri nu doar că-i vor conecta pe oameni între ei, dar îi vor conecta şi pe oameni cu maşinile şi, crucial, maşinile cu maşinile.

Wiener a întrezărit o lume în care informaţia va putea fi separată de mediul său. Oamenii sau maşinile ar putea comunica tiparele lor pe distanţe uriaşe şi le vor putea utiliza pentru a produce obiecte noi în punctele de destinaţie, fără să „mişte o… particulă de materie de la un capăt al liniei la celălalt", o viziune realizată acum în lumea noastră de imprimante 3D în reţea.

Wiener şi-a imaginat, de asemenea, şi cicluri de *feedback* de la o maşină la alta, care vor sta la baza unor avansuri uriaşe în automatizare, chiar şi pentru sarcini care s-au bazat înainte pe judecata umană. „Maşina nu are preferinţe când trebuie să aleagă între munca fizică şi munca de birou", a observat el.

Cu toate acestea, multe dintre argumentele centrale din *The Human Use of Human Beings* par să fie mai apropiate de secolul al XIX-lea decât de secolul al XXI-lea. În special, deşi Wiener a făcut referire în toată lucrarea sa la ceea ce era pe atunci noua teorie a informaţiei emisă de Claude Shannon, se pare că nu a îmbrăţişat cu totul noţiunea lui Shannon de informaţie care constă în biţi ireductibili şi fără înţeles propriu. Din vremea lui Wiener, teoria lui Shannon a stat la baza avansurilor recente în domeniile „Big Data" şi „învăţare automată", ceea ce face cu atât mai interesantă revizitarea imaginaţiei cibernetice a lui Wiener. Cum ar putea fi diferită inteligenţa artificială de mâine dacă practicanţii ar fi să reinvestească în viziunea călăuzitoare a lui Wiener despre „informaţie"?

Când Wiener scria *The Human Use of Human Beingsy* experienţele

sale în cercetările legate de război şi de ceea ce i se păreau ambiguităţile morale ale vieţii intelectuale în cadrul complexului militar industrial erau încă proaspete. Cu doar câţiva ani înainte, anunţase în paginile revistei *Atlantic Monthly* că nu va „publica lucrări viitoare care ar putea cauza rău în mâinile militariştilor iresponsabili”[53]. A rămas ambivalent cu privire la puterea transformatoare a noilor tehnologii, nelăsându-se nici dus de valul tendinţelor la modă şi nici de utopismul digital al savanţilor care i-au urmat.

„Progresul impune nu doar posibilităţi noi pentru viitor, dar şi restricţii noi”, scria el în *The Human Use of Human Beings*. Era îngrijorat atât de restricţiile umane, cât şi de cele tehnologice, în special restricţiile din timpul Războiului Rece care ameninţau fluxul de informaţii atât de important pentru sistemele cibernetice: „Sub elanul senatorului [Joseph] McCarthy şi al imitatorilor săi, clasificarea oarbă şi excesivă a informaţiilor militare” îi determina pe liderii politici din Statele Unite să adopte o „mentalitate atât de discretă încât era egalată doar de perioada Veneţiei renascentiste”. Wiener, spunând ce au spus şi mulţi veterani ai Proiectului Manhattan, argumenta că obsesia confidenţialităţii după război – în special în privinţa armelor nucleare – îşi avea rădăcinile în neînţelegerea procesului ştiinţific. Singurul secret veritabil în ceea ce priveşte producţia armelor nucleare, scria el, era dacă astfel de bombe pot fi construite cu adevărat. După dezvăluirea acestui secret, cu bombardarea oraşelor Hirosima şi Nagasaki, oricât de mult ar încerca statul să acopere secretul, oamenii nu ar putea fi opriţi să gândească etapele urmate de acei cercetători din Proiectul Manhattan. În memorabilele cuvinte ale lui Wiener, „nu există linia Maginot pentru creier”.

Pentru a-şi demonstra spusele, Wiener a împrumutat ideile proaspete ale lui Shannon despre teoria informaţiei. În 1948, Shannon, matematician şi inginer care lucra la Bell Labs, a publicat două articole lungi în revista *Bell Sistem Tehnical Journal*. Prezentând aceste idei noi pentru mai multe categorii de cititori în 1949, matematicianul Warren Weaver a explicat că, în formularea lui Shannon, „cuvântul *informaţie*… este utilizat cu un sens special care nu trebuie confundat cu utilizarea sa comună. În mod special, *informaţia* nu trebuie confundată cu sensul”[54]. Lingviştii şi poeţii se preocupă de aspectele „semantice” ale comunicării, continuă Weaver, dar nu şi inginerii ca Shannon. Mai degrabă, „acest cuvânt, *informaţie*, în teoriile comunicării se referă nu atât de mult la *ceea ce spui,* ci mai mult la *ceea ce ai putea spunea*. În cuvintele acum renumite ale lui Shannon, conţinutul de informaţie al unui şir de simboluri era dat de logaritmul numărului de simboluri posibile din care a fost ales un şir

dat. Observaţiile-cheie ale lui Shannon erau că informaţia unui mesaj era asemănătoare entropiei unui gaz: o măsură a dezordinii din sistem.

Wiener a împrumutat această observaţie când a scris *The Human Use of Human Beings*. Dacă informaţia era asemănătoare entropiei, atunci nu putea fi conservată sau controlată. Fizicienii secolului al XIX-lea au dovedit că energia totală a unui sistem fizic trebuie să rămână întotdeauna constantă, un echilibru perfect între începutul şi sfârşitul unui proces. Lucrurile nu stau la fel şi cu entropia, care ar creşte inexorabil în timp, un imperativ care a ajuns să fie cunoscut drept a doua legii a termodinamicii. Din această puternică diferenţă – energia este conservată, iar entropia trebuie să crească – reieşeau consecinţe cosmice enorme. Timpul trebuie să curgă înainte; viitorul nu poate fi identic cu trecutul. Universul se poate îndrepta către o „moarte termică", într-un timp foarte îndepărtat, când stocul de energie va fi dispersat uniform, ajungând într-o stare de entropie maximă, după care nu mai poate apărea nicio schimbare ulterioară.

Dacă informaţia *ca* entropie *per* să nu ar putea fi conservată, Wiener a conchis că este o nebunie ca liderii militari să încerce să acumuleze *Jknow-how-xA* ştiinţific al naţiunii în biblioteci şi laboratoare statice". Într-adevăr, „indiferent de volumul de cercetări ştiinţifice, înregistrate cu atenţie în cărţi şi lucrări de specialitate şi introduse apoi în bibliotecile noastre cu etichete de secretizare, nu ne vom putea proteja într-o lume în care nivelul efectiv de informaţie este în continuă creştere". Orice astfel de efort de secretizare, clasificare sau colecţionare a acestor informaţii era sortit pieirii, spunea Wiener, la fel cum escrocheriile pungaşilor pentru maşini cu mişcare perpetuă nu făceau faţă celei de-a doua legi a termodinamicii.

Wiener a criticat „ortodoxia" americană a fundamentalismului pieţelor libere aproape în acelaşi fel. Pentru majoritatea americanilor, „problemele privind informaţia vor fi evaluate conform unui criteriu american standard: un lucru este valoros ca bun de consum pentru ce introduce pe piaţa deschisă". Într-adevăr, „soarta informaţiei în lumea tipic americană urma să devină ceva ce poate fi cumpărat sau vândut"; majoritatea oamenilor, observa el, „nu pot concepe ideea ca informaţia să nu aibă un proprietar". Wiener considera această idee la fel de greşită precum clasificările militare exagerate. Din nou, invoca observaţiile lui Shannon: „deoarece informaţiile şi entropia nu sunt conservate", acestea sunt „la fel de nepotrivite pentru a deveni bunuri de consum".

Informaţia nu poate fi conservată – până aici, totul este clar. Dar se referea Wiener la „informaţia" lui Shannon? Partea dificilă din argumentul lui Shannon, după cum observa Wiener, era să distingă

sensul colocvial al „informaţiei”, ca mesaj cu sens, de noţiunea abstractă, rarefiată de şiruri de simboluri aranjate cu o anumită probabilitate şi selectat dintr-un univers enorm de vorbe fără noimă. Pentru Shannon, „informaţia” putea fi cuantificată, deoarece unitatea sa fundamentală, bitul, era o unitate de transmitere, şi nu una de sens.

Când Wiener a caracterizat „informaţia” în cartea *The Human Use of Human Beings*, pe de altă parte, a făcut referire de mai multe ori la un sens clasic, umanist al termenului. „O bucată de informaţie”, scria el – atenţie, nu un „bit” de informaţie –, „pentru a putea contribui la informaţia generală a comunităţii, trebuie să spună ceva substanţial diferit de stocul de informaţii comun anterior al comunităţii”. Acesta este motivul pentru care „şcolarii nu-l apreciază pe Shakespeare”, concluziona el: stihurile bardului pot porni brusc din fluxuri aleatorii de biţi, dar au devenit cu toate acestea foarte cunoscute publicului decodificator de sens şi „absorbite în clişeele superficiale ale vremii”.

Cel puţin, conţinutul informaţional al lui Shakespeare a părut cândva proaspăt. În anii de după *boom-ul* postbelic, Wiener se văita că „volumul enorm de comunicare *per capita* – de la ziare şi filme până la radio, televiziune şi cărţi – a generat mediocritate, o revenire informaţională la medie. „Din ce în ce mai mult, trebuie să acceptăm un produs standardizat inofensiv şi nesemnificativ care, la fel ca pâinea albă de la brutărie, este făcut mai mult pentru proprietăţile sale de păstrare şi vandabilitate decât pentru valoarea nutritivă”. „Să ne ferească Dumnezeu”, exclamă el, „de primele romane scrise, pentru că un tânăr doreşte prestigiul de a fi romancier, şi nu că ar avea ceva de spus! Să ne ferească Dumnezeu şi de lucrările matematice care sunt corecte şi elegante, dar fără trup şi spirit”. Felul în care privea Wiener „informaţia” suna mai mult a Matthew Arnold în 1869[55] decât a Claude Shannon în 1948 – mai mult „trup şi spirit” decât „bit”. Şi Wiener împărtăşea viziunea romantică a lui Arnold a „producătorului de conţinut”: „la drept vorbind, artistul, scriitorul şi omul de ştiinţă ar trebui să fie mişcat de un impuls atât de irezistibil de a crea încât, chiar dacă nu ar fi plătit pentru munca depusă, să fie dispus să plătească el însuşi pentru şansa de a o face”. *L’artpour Vart*, acel strigăt al secolului al XIX-lea: artiştii trebuie să sufere pentru munca lor; căutarea expresiei cu sens trebuie să bată întotdeauna câştigul.

Pentru Wiener, aceasta era măsura corectă a „informaţiei”: trup, spirit, aspiraţie, expresie. Însă, pentru a argumenta împotriva transformării acesteia în bun de consum, Wiener s-a întors încă o dată la matematica lui Shannon care privea „informaţia ca entropie”.

Sărim până în zilele noastre. În multe feluri, s-a dovedit că Wiener avea dreptate. Viziunea sa privind ciclurile de *feedback* în reţea guvernate de comunicaţii de la maşină la maşină a devenit o

caracteristică obişnuită a vieţii de zi cu zi. În plus, încă din cele mai vechi zile ale Epocii Internetului, pirateria digitală a desfiinţat imaginea conform căreia „informaţia” – sub formă de melodii, filme, cărţi sau cod – poate rămâne închisă. Pune un abonament cu *paywall* aici şi conţinutul va fi difuzat acolo, atât de multă entropie informaţională încât nu poate fi conservată.

Pe de altă parte, corporaţii multinaţionale enorme – câteva dintre cele mai mari şi mai profitabile din lume – dezaprobă acum la ordinea zilei punctele de vedere controversate care susţin că „informaţia” nu poate fi stocată sau monetizată. Ironic, „informaţia” pe care o comercializează ele este mai aproape de definiţia lui Shannon decât de cea a lui Wiener, neţinând cont de dovezile matematice ale lui Shannon.

Deşi Google Books ajută sute de mii de cărţi de literatură să circule gratuit, Google în sine – la fel ca Facebook, Amazon, Twitter şi multe alte companii care le imită – a confiscat o formă mai simplă de „informaţie” şi a exploatat-o pentru un profit extraordinar. *Petabytes* de informaţii de tip Shannon – un şir aparent fără valoare de clicuri, aprecieri şi *retweet-uri* colectate de la fiecare persoană care a atins vreodată un computer conectat în reţea – sunt cernute prin algoritmi brevetaţi de „învăţare automată” pentru a microţinti totul, de la reclamele pe care le vedem până la ştirile (false sau nu) pe care le întâlnim când navigăm pe internet.

La începutul anilor cincizeci, Wiener a propus ca cercetătorii să studieze structurile şi limitările furnicilor – faţă de ale oamenilor – pentru ca maşinile să obţină într-o bună zi „aproape indefinita extindere intelectuală” pe care oamenii o pot atinge (şi insectele nu). A găsit consolare în ideea că maşinile ar putea ajunge să ne domine doar „în ultimele stadii ale entropiei crescânde”, când „diferenţele statistice dintre indivizi vor fi nule”. Algoritmii de minare a datelor de astăzi dau peste cap teoria lui Wiener. Produc profit prin exploatarea creierului nostru reptilian, şi nu prin imitarea cortexurilor noastre cerebrale, culegând informaţii din toate fluxurile noastre de clicuri făcute noaptea târziu, rătăcite pe bloguri şi căutătoare de plăceri – valorificând exact micile şi rezidualele „diferenţe statistice dintre indivizi”.

Cu siguranţă, câteva realizări recente din domeniul inteligenţei artificiale au fost remarcabil de impresionante. Computerele pot produce acum opere de artă vizuală şi compoziţii muzicale asemănătoare celor ale maeştrilor recunoscuţi şi creând exact tipul de „informaţie” pe care Wiener o valoriza cel mai mult. Dar, de departe, cel mai mare impact asupra societăţii până astăzi a venit de la colectarea şi manipularea informaţiilor de tip Shannon, care au dat o nouă formă obiceiurilor noastre de cumpărare, participării politice,

relaţiilor personale, aşteptărilor privind confidenţialitatea şi multe altele.

În ce ar putea evolua „învăţarea profundă" dacă moneda fundamentală devine „informaţia" aşa cum a definit-o Wiener? Cum s-ar schimba domeniul dacă ar fi reanimat de convingerile morale profunde ale lui Wiener, informate fiind de îngrijorările sale vizionare cu privire la militarismul frenetic, setea nestăvilită de profit a corporaţiilor, caracteristicile autolimitante ale secretelor şi reducerea expresiei umane la bunuri de consum interschimbabile? Poate că „învăţarea profundă" ar putea deveni cultivarea informaţiei semnificative în locul urmăririi neîncetate a biţilor potenţi, chiar dacă sunt lipsiţi de sens.

Capitolul 16
Scalarea

Neil Gershenfeld

Neil Gershenfeld este fizician şi director al Centrului pentru Biţi şi Atomi de la MIT. Este autorul cărţilor FABy coautorul (împreună cu Alan Gershenfeld şi Joel Cutcher-Gershenfeld) cărţii Designing Reality *şi fondatorul reţelei globale* Fab Lab.

În sus-menţionată discuţie din Connecticut despre The Human Use of Human Beings, Neil Gerhenfeld a oferit o perspectivă proaspătă, spunând că nu i-a plăcut deloc cartea. Observaţia sa a fost întâmpinată cu râsete generale – la fel şi observaţia sa că informatica a fost unul dintre cele mai rele lucruri care s-a întâmplat computerelor sau ştiinţei. Controversa sa generală a fost că Wiener a ratat implicaţiile revoluţiei digitale care avea loc în jurul său, deşi unii ar spune că această acuzaţie nu poate fi făcută despre o persoană de la cel mai jos nivel şi căreia îi lipseşte clarviziunea.

„Coada care scutură câinele vieţii mele", ne-a spus el, „a fost Fab Labs şi mişcarea ei de creare, şi, când Wiener vorbeşte despre ameninţarea automatizării, scapă din vedere reciproca, şi anume că accesul la mijloacele de automatizare îi poate împuternici pe oameni, iar, în Fab Labs, colţul în care am fost implicat, acest lucru este exponenţial".

În 2003, i-am făcut o vizită lui Neil la MIT, unde conduce Centrul pentru Biţi şi Atomi. Câteva ore mai târziu, am ieşit din ceea ce fusese o afişare exuberantă de lucruri foarte ciudate. Mi-a arătat activitatea unuia dintre studenţii săi la popularul curs pe care-l ţinea despre crearea rapidă a prototipurilor (Cum să faci

aproape orice), un sculptor fără studii de inginerie care a făcut un spaţiu personal portabil de ţipat care îţi salvează ţipetele şi le poţi asculta mai târziu. Alt student de la cursul său a făcut un browser web care permite papagalilor să navigheze pe internet. Neil însuşi făcea cercetări fundamentale pe harta care duce la acel vis science-fiction, „replicatorul universal". A fost o vizită pentru care am avut nevoie de câţiva ani s-o înţeleg.

Neil administrează reţeaua globală Fab Labs – sisteme de producţie la scară mică, având la bază tehnologii digitale, ce oferă oamenilor mijloacele necesare pentru a construi ce vor ei. În calitate de guru al acestei mişcări, care uneşte comunicarea digitală şi tehnica de calcul cu procesul de producţie, uneori se simte în afara dezbaterii aprinse privind siguranţa inteligenţei artificiale. „Abilitatea mea de a face cercetări se bazează pe instrumente care-mi augmentează capacităţile", spune el. „A întreba dacă sunt inteligente sau nu este la fel de util ca a întreba cum ştiu că exist – distractiv din punct de vedere filosofic, dar nu poate fi testat empiric". Ce îl interesează pe el este „cum interacţionează biţii şi atomii – graniţa dintre digital şi fizic. Din punct de vedere ştiinţific, este cel mai interesant lucru pe care îl ştiu".

Discuţiile pe tema inteligenţei artificiale au fost dubios de anistorice. Ar putea fi descrise mai bine ca maniaco-depresive; în funcţie de cum numărăm, ne aflăm acum în al cincilea ciclu de expansiune şi contracţie. Aceste pendulări maschează continuitatea din progresul subiacent şi implicaţiile punctului către care se îndreaptă.

Ciclurile au venit în valuri de aproximativ zece ani. Mai întâi, au fost computerele *mainframe*, care, prin simpla lor existenţă, urmau să automatizeze munca. Ideea s-a izbit de zidul realităţii când s-a constat că este greu să scrii programe care să efectueze sarcini uşor de realizat pentru oameni. Apoi au venit experţii de sisteme, care urmau să codifice şi apoi să-i înlocuiască pe experţii în cunoştinţe. Aceştia au dat de greu în asamblarea cunoştinţelor şi raţionării cu privire la cazurile care nu erau deja acoperite. Percepţionii căutau să evite aceste probleme modelând modul în care învaţă creierul, dar nu au făcut o treabă prea bună. Percepţionii cu mai multe niveluri puteau să gestioneze probleme-test care declanşau acele reţele simple, dar demonstraţiile lor nu s-au descurcat cu probleme nestructurate, reale. Acum ne aflăm în era învăţării profunde, care a reuşit în privinţa multor promisiuni timpurii ale inteligenţei artificiale, dar într-un mod considerat greu de înţeles, cu consecinţe care variază de la ameninţări intelectuale până la ameninţări existenţiale.

Fiecare dintre aceste etape era anunţată ca un avans revoluţionar

asupra limitărilor predecesorilor, însă toate trebuiau să facă practic acelaşi lucru: să facă inferenţe din observaţii. Modul cum relaţionează aceste abordări poate fi înţeles din cum pot fi scalate – mai exact, cum depinde performanţa lor de dificultatea problemei pe care încearcă să o rezolve. Atât întrerupătorul becului, cât şi maşina autonomă trebuie să stabilească intenţiile operatorilor, dar primul are doar două opţiuni din care să aleagă, în timp ce maşina are mult mai multe. Etapele boomului inteligenţei artificiale au început cu exemple promiţătoare în domenii limitate; fazele de extindere au venit cu nereuşita demonstraţiei de a gestiona complexitatea problemelor practice mai puţin structurate.

Mai puţin vizibil este progresul continuu pe care l-am făcut în domeniul scalării. Acest progres are la bază distincţia tehnologică dintre funcţiile liniare şi cele exponenţiale – o distincţie care devenea evidentă la începuturile inteligenţei artificiale, dar cu implicaţii pentru inteligenţa artificială care nu au fost apreciate decât mulţi ani mai târziu.

Într-unul dintre documentele fondatoare despre studierea maşinilor inteligente, *The Human Use of Human Beings*, Norbert Wiener face o treabă remarcabilă în identificarea celor mai semnificative tendinţe care urmau să ia naştere după ce a scris cartea, observându-i în acelaşi timp şi pe oamenii responsabili de ele, apoi nereuşind niciodată să recunoască de ce munca acestora se va dovedi atât de importantă. Wiener este considerat a fi inventatorul domeniului ciberneticii; nu am înţeles niciodată ce-i asta, dar ceea ce lipseşte din carte este tocmai nucleul progresului inteligenţei artificiale. Această istorie contează, pentru că ecourile sale au continuat până în prezent.

Claude Shannon apare şi el în carte, într-un rol secundar, în contextul ideilor sale despre perspectivele unui computer capabil să joace şah. Shannon făcea ceva mult mai semnificativ decât să speculeze la vremea aceea: turna fundaţia unei revoluţii digitale. Absolvent al Institutului Tehnologic din Massachusetts, a lucrat pentru Vannevar Bush la Analizorul Diferenţial. Acesta a fost unul dintre ultimele mari computere analogice, o cameră plină de roţi şi axuri. Frustrarea lui Shannon în ceea ce priveşte dificultatea rezolvării problemelor în acest mod l-a făcut să scrie în 1937 probabil cea mai bună dizertaţie de masterat din toate timpurile. În aceasta, a arătat cum circuitele electrice pot fi proiectate să evalueze expresii logice arbitrare, introducând baza pentru logica digitală universală.

După MIT, Shannon a studiat comunicaţiile la Bell Labs. Apelurile telefonice analogice se deteriorau proporţional cu distanţa; cu cât ajungeau mai departe, cu atât sunau mai rău. Decât să continuăm să le îmbunătăţim incremental, Shannon a arătat în 1948 că, prin comunicare cu simboluri mai degrabă decât cu cantităţi continue,

comportamentul este foarte diferit. Convertirea formelor de undă ale limbajului la valorile binare de 1 şi o este un exemplu, dar multe alte seturi de simboluri pot fi (şi sunt) utilizate în comunicări digitale. Ce contează nu sunt simbolurile particulare, ci mai degrabă capacitatea de a detecta şi corecta erori. Shannon a descoperit că, dacă zgomotul este peste o anumită valoare (care depinde de designul sistemului), atunci cu siguranţă vor apărea erori. Dar, dacă zgomotul este sub un anumit nivel, atunci o creştere liniară în resursele fizice ce reprezintă simbolul determină o scădere exponenţială a probabilităţii de a face o greşeală în receptarea corectă a simbolului. Această relaţie a fost prima pentru ceea ce numim în prezent o teoremă a valorii de prag.

O astfel de scalare cade atât de rapid, încât probabilitatea unei erori poate fi atât de mică încât efectiv să nu apară niciodată. Fiecare simbol trimis multiplică, şi nu adaugă certitudinea, astfel încât probabilitatea unei greşeli să scadă de la 0,1 la 0,01 la 0,001 şi aşa mai departe. Această scădere exponenţială în erorile de comunicare a făcut posibilă o creştere exponenţială a capacităţii reţelelor de comunicare. Şi acest lucru a rezolvat în cele din urmă problema originii cunoştinţelor unui sistem de inteligenţă artificială.

Mulţi ani, cel mai rapid mod de a îmbunătăţi o tehnică de calcul era să nu faci nimic – doar să aştepţi computerele să devină mai rapide. În acelaşi mod, au fost ani întregi de proiecte de inteligenţă artificială care urmăreau să acumuleze cunoştinţe de zi cu zi introducând laborios bucăţi de informaţie. Acest model nu se putea scala; putea progresa doar proporţional cu numărul de oameni care introduceau datele. Dar când apelurile telefonice, articolele de ziar şi mesajele de e-mail s-au mutat pe internet, toată lumea care făcea oricare dintre cele menţionate a devenit automat un generator de date. Rezultatul a fost o rată de acumulare a cunoştinţelor exponenţială, şi nu liniară.

Şi John von Neumann are un rol secundar în *The Human Use of Human Beings* în ceea ce priveşte teoria jocurilor. Ce a ratat Wiener aici este rolul seminal al lui Neumann în digitalizarea tehnicii de calcul. În timp ce comunicările analogice se degradau proporţional cu distanţa, tehnica de calcul analogică (de exemplu, Analizorul Diferenţial) se degrada în timp, acumulând erori pe măsură ce progresa. Von Neumann a prezentat în 1952 un rezultat ce corespundea cu al lui Shannon în ceea ce priveşte tehnica de calcul (s-au întâlnit la Institutul de Studii Avansate din Princeton), arătând că este posibilă o tehnică de calcul fiabilă cu un dispozitiv de calcul nefiabil utilizând simboluri în locul cantităţilor continue. Acesta era, din nou, un argument pentru scalare cu o creştere liniară a resurselor fizice ce reprezintă simbolul şi rezultând într-o reducere exponenţială a ratei erorilor, atât timp cât zgomotul era sub un anumit prag. Asta-i

ceea ce face posibilă integrarea unui miliard de tranzistori pe un cip de computer, cu ultimul fiind la fel de util ca primul. Această relaţie a condus la o creştere exponenţială în performanţa de calcul, ceea ce a rezolvat o a doua problemă a inteligenţei artificiale: prelucrarea unui volum de date aflat într-o creştere exponenţială.

A treia problemă pe care a rezolvat-o scalarea pentru inteligenţa artificială a fost structurarea regulilor de raţionare fără a fi nevoie de angajarea unui programator pentru fiecare problemă. Wiener a recunoscut rolul feedbackrului în învăţarea automată, dar a ratat rolul-cheie al reprezentării. Nu este posibil să stochezi toate imaginile posibile într-o maşină autonomă sau toate sunetele posibile într-un computer conversaţional; trebuie să poată generaliza din experienţă. Partea „profundă" din învăţarea profundă nu se referă la profunzimea (sperată a) observaţiilor, ci la profunzimea straturilor de reţele matematice utilizate pentru a face predicţii. S-a dovedit că o creştere liniară în complexitatea reţelei conduce la o creştere exponenţială în puterea de expresie a reţelei.

Dacă-ţi pierzi cheile într-o cameră, le poţi căuta. Dacă nu eşti sigur în ce cameră se află, trebuie să cauţi în toate camerele din clădire. Dacă nu eşti sigur în ce clădire se află, trebuie să cauţi în toate clădirile din oraş. Dacă nu ştii sigur în ce oraş te afli, trebuie să cauţi în toate camerele din toate clădirile din toate oraşele. În inteligenţa artificială, găsirea cheilor corespunde unor lucruri precum o maşină care urmează în siguranţă un drum sau un computer care interpretează corect o comandă vorbită, iar camerele, clădirile şi oraşele corespund tuturor opţiunilor care trebuie luate în considerare. Acesta este numit blestemul dimensionalităţii.

Soluţia blestemului dimensionalităţii a venit din folosirea informaţiilor despre problemă pentru a restricţiona căutarea. Algoritmii de căutare în sine nu sunt noi, dar, când sunt aplicaţi la o reţea cu învăţare profundă, construiesc adaptiv reprezentări pentru locul în care ar trebui să caute. Preţul acestui lucru este că nu mai este posibil să rezolvi exact pentru cel mai bun răspuns la o problemă, dar, de obicei, este necesar doar un răspuns care să fie suficient de bun.

Luate împreună, nu ar trebui să fie surprinzător că aceste legi de scalare au permis maşinilor să devină efectiv la fel de capabile precum stadiile corespunzătoare ale complexităţii biologice. Reţelele neuronale au pornit iniţial cu obiectivul de a modela cum funcţionează creierul. Acel obiectiv a fost abandonat, deoarece au evoluat în abstracţii matematice care nu mai aveau nicio legătură cu modul în care funcţionează neuronii. Dar acum a apărut un soi de convergenţă care poate fi considerată o biologie cu inginerie normală, şi nu inginerie inversă, deoarece rezultatele învăţării profunde oglindesc nivelurile şi regiunile creierului.

Unul dintre cele mai dificile proiecte de cercetare pe care le-am gestionat a pus laolaltă ceea ce numim acum cercetători de date şi pionieri ai inteligenţei artificiale. A fost o experienţă mizerabilă de mutare a barelor porţii. Cât timp cercetătorii de date progresau în rezolvarea problemelor vechi ridicate de inteligenţa artificială, se considera că nu contează, pentru că nu era realizat prin salturi în înţelegerea soluţiilor. Care este valoarea unui computer care joacă şah dacă nu poţi să explici cum joacă şah?

Răspunsul, bineînţeles, este că poate juca şah. Există cercetări emergente interesante care aplică inteligenţa artificială pe inteligenţa artificială – adică instruiesc reţele ca să explice cum funcţionează. Dar atât creierele, cât şi cipurile de computer sunt greu de înţeles dacă le priveşti mecanismele; se interpretează uşor doar prin observarea interfeţelor interne. Ajungem să avem încredere (sau nu) în creiere şi cipuri de computer în baza unor experienţe şi a unor teste, şi nu în funcţie de explicaţiile pe care le primim pentru cum funcţionează.

Multe branşe de proiectare fac o tranziţie de la ceea ce este numit design imperativ la ceea ce este numit design declarativ sau generativ. Asta înseamnă că, în loc de a proiecta explicit un sistem cu instrumente ca fişierele CAD, scheme de circuite şi cod de computer, descrii ce vrei să facă sistemul şi apoi este efectuată o căutare automată pentru designurile care satisfac obiectivele şi restricţiile tale. Această abordare devine necesară pe măsură ce complexitatea designului depăşeşte ceea ce poate fi înţeles de un proiectant uman. Deşi sună ca un risc, înţelegerea omenească vine cu propriile sale limitări; proiectarea în inginerie este presărată cu ceea ce au părut a fi observaţii bune, dar au avut consecinţe negative. Designul declarativ se bazează pe toate progresele din domeniul inteligenţei artificiale, plus fidelitatea din ce în ce mai bună a simulărilor care testează virtual designurile.

Mama tuturor problemelor de design este cea care ne-a creat pe noi. Modul în care suntem proiectaţi rezidă în una dintre cele mai vechi şi bine conservate părţi ale genomului, numită genele Hox. Acestea sunt gene care reglează genele, în ceea ce se numesc programe de dezvoltare. Nimic din genomul nostru nu memorează structura corpului nostru; genomul memorează, mai degrabă, o serie de etape care urmăresc acel rezultat în corpul nostru.

Aceasta este o paralelă exactă pentru cum se fac căutările în inteligenţa artificială. Există prea multe planuri posibile pentru corpuri prin care să caute şi cele mai multe modificări ar fi fie neimportante, fie fatale. Genele Hox sunt o reprezentare a unui loc productiv pentru căutare evoluţionistă. Sunt un fel de inteligenţă naturală la nivel molecular.

Inteligenţa artificială are o problemă minte-corp în sensul că nu

are corp. Mare parte a muncii inteligenţei artificiale este făcută în *cloud*, ce rulează pe maşini virtuale în centre de computere unde sunt direcţionate datele. Propria noastră inteligenţă este rezultatul unui algoritm de căutare (evoluţia) care ne-a putut schimba forma fizică, precum şi programarea – acestea sunt legate inextricabil. Dacă istoria inteligenţei artificiale poate fi înţeleasă ca rezultatul legilor de scalare, mai degrabă decât o succesiune de mode, atunci viitorul său poate fi văzut în acelaşi mod. Ce este acum digitalizat, după comunicaţie şi tehnică de calcul, este producţia, aducerea capacităţii de programare a biţilor în lumea atomilor. Digitalizând nu doar designurile, ci şi construcţia materialelor, aceleaşi lecţii pe care le-am învăţat de la von Neumann şi Shannon se aplică şi complexităţii cu creştere exponenţială a proceselor de producţie.

Am definit materialele digitale ca fiind cele construite dintr-un set discret de părţi reversibil combinat cu un set discret de poziţii şi orientări relative. Aceste atribute permit determinarea geometriei globale din restricţii locale, detectarea şi corectarea erorilor de asamblare, combinarea materialelor eterogene şi demontarea structurilor când nu mai sunt necesare, în loc de distrugerea lor. Aminoacizii ce reprezintă temelia vieţii şi piesele Lego ce reprezintă fundaţia jocului împărtăşesc aceste proprietăţi.

Ce este interesant cu privire la aminoacizi este că nu sunt interesanţi. Au atribute care sunt tipice, dar nu neobişnuite, cum ar fi să poată atrage sau respinge apa. Dar nu mai mult de douăzeci de tipuri dintre ei sunt suficiente ca să-ţi dea viaţă. În anumite feluri, douăzeci şi ceva de tipuri de piese de material digital – conductoare, izolante, rigide, flexibile, magnetice etc. – sunt suficiente pentru a asambla un interval de funcţii care participă la fabricarea tehnologiilor moderne, precum roboţii şi computerele.

Legătura dintre tehnică de calcul şi fabricaţie a fost prefigurată chiar de acei pionieri a căror muncă este fundaţia edificiului tehnicii de calcul. Wiener a făcut aluzie la ea corelând transportul de material cu transferul de mesaje. John von Neumann este creditat ca inventator al arhitecturii computerelor moderne, subiect despre care, de fapt, a scris foarte puţin; ultimul lucru pe care l-a studiat şi despre care a scris frumos şi pe îndelete au fost sistemele cu reproducere automată. Ca o abstractizare a vieţii, a modelat o maşină care poate comunica o tehnică de calcul ce se construieşte pe sine. Şi ultimul lucru pe care l-a studiat Alan Turing, care este creditat cu cadrul teoretic al informaticii, a fost cum pot instrucţiunile din gene să dea naştere formelor fizice. Aceste întrebări vizează o temă ce lipseşte din educaţia informatică obişnuită: configuraţia fizică a unei tehnici de calcul.

Von Neumann şi Turing şi-au pus întrebările ca studii teoretice,

pentru că erau dincolo de posibilităţile tehnologiei din vremea lor şi nu puteau fi puse în practică. Dar, odată cu convergenţa comunicaţiilor şi tehnicii de calcul cu producţia, aceste investigaţii devin acum accesibile experimental. În laboratorul meu ne concentrăm pe asamblarea unui robot de asamblare care se poate asambla pe sine din părţile componente, dar colaborăm şi cu alte companii pentru a dezvolta celule sintetice.

Ideea automatelor care se pot reproduce automat şi fizic are un potenţial mult mai înspăimântător decât temerile legate de inteligenţa artificială scăpată de sub control, pentru că mută inteligenţa în afară, unde trăim noi. Poate fi o traiectorie care ne va duce la stăpânii robotici Skynet din filmul *Terminator*. Dar există şi idei mai luminoase, pentru că abilitatea de a programa atomi la fel de bine cum programăm biţii permite designurilor să fie împărtăşite global, producând în acelaşi timp local energie, alimente şi adăpost – toate acestea sunt emergente sub formă de aplicaţii noi şi interesante al fabricaţiei digitale. Wiener îşi făcea probleme în legătură cu viitorul muncii, dar nu a pus la îndoială presupunerile implicite cu privire la natura muncii care sunt provocate atunci când consumul poate fi înlocuit prin creaţie.

Istoria ne învaţă că nu prevalează nici scenariile utopice, nici cele distopice; în general, orbecăim undeva între. Dar istoria ne învaţă şi că nu trebuie să aşteptăm istoria. Gordon Moore, în 1965, a putut utiliza cinci ani de dublare a specificaţiilor circuitelor integrate pentru a proiecta ceea ce s-au dovedit a fi cincizeci de ani de îmbunătăţiri exponenţiale în tehnologiile digitale. Am petrecut mulţi din acei ani răspunzând la în loc să anticipăm, implicaţiile sale. Avem mai multe date disponibile acum decât avea Gordon Moore pentru a proiecta cincizeci de ani de dublare a performanţelor producţiei digitale. Înarmaţi cu beneficiile retrospectivei, ar trebui să fie posibil să evităm excesele tehnicii de calcul digitale şi comunicaţiilor de data aceasta şi, pentru început, să ne ocupăm de problemele de acces şi alfabetizare.

Dacă mişcarea celor ce produc este vestitorul unei a treia revoluţii digitale, succesul unei inteligenţe artificiale în îndeplinirea multor obiective iniţiale poate fi văzut ca realizarea principală a primelor două revoluţii digitale. Deşi producţia de maşini şi modul în care gândesc maşinile par să fie două tendinţe separate, în viitor depind una de cealaltă. Aceleaşi tendinţe de scalare care au făcut posibilă inteligenţa artificială sugerează că mania curentă este o fază trecătoare, care va fi urmată de ceva şi mai important: îmbinarea inteligenţei artificiale cu inteligenţa naturală.

A fost un avans când atomii au format molecule, moleculele au format organite celulare, organitele au format celule, celulele au format organe, organele au format organisme, organismele au format

familii, familiile au format societăţi şi societăţile au format civilizaţii. Acest ciclu evoluţionist la scară mare poate fi închis acum cu atomi care aranjează biţi aranjatori de atomi.

Capitolul 17
Primele inteligenţe automate

W. Daniel Hills

W. Daniel „Danny" Hillis este inventator, antreprenor şi informatician, profesor Judecător Widney" de inginerie şi medicină la USC şi autorul cărţii The Pattern on the Stone: The Simple Ideas That Make Computers Work.

Cât timp Danny Hillis a fost student la MIT, a construit un computer din jucării de lemn. Are aproximativ 10.000 de piese de lemn, joacă X şi o şi nu pierde niciodată; acum se află la Muzeul de Istorie a Computerelor din Mountain View, California.

După ce a absolvit şi lucra la Laboratorul de inteligenţă artificială şi informatică de la MIT, la începutul anilor 1980, Danny a conceput un computer masiv care funcţiona în paralel, cu 64.000 de procesoare. L-a numit Maşina Conexiunilor şi a înfiinţat ceea ce a fost probabil prima companie de inteligenţă artificială – Thinking Machines Corporation – pentru a-l produce şi a-l comercializa. Aceasta, în ciuda unui prânz pe care l-a luat împreună cu Richard Feynman în timpul căruia renumitul fizician a remarcat, „asta e de-a dreptul cea mai nătângă idee pe care am auzit-o vreodată". Poate „în ciuda" este cuvântul nepotrivit, de vreme ce Feynman avea o predilecţie bine-cunoscută de a se juca cu ideile nătânge. La eveniment, s-a prezentat în ziua înfiinţării companiei şi a rămas în comitetul de conducere, pentru activităţi pe timpul verii şi sarcini speciale, aducând astfel contribuţii importante la activitatea sa.

Danny a înfiinţat de atunci mai multe companii de tehnologie, dintre care cea mai nouă este Applied Invention, care face parteneriate cu întreprinderi comerciale pentru a dezvolta soluţii tehnologice pentru cele mai dificile probleme pe care le au. Deţine sute de brevete în SUA, care acoperă computerele în paralel, interfeţele tactile, matricele de discuri, metode de împiedicare a falsurilor şi multe dispozitive mecanice şi electrice. Imaginaţia sa este, aparent, fără limite şi schiţează aici câteva scenarii posibile ce vor lua naştere din urmărirea unei inteligenţe artificiale din ce în ce mai bune.

„Maşinile noastre gânditoare sunt mai mult decât nişte simple metafore", spune el. „Întrebarea nu este dacă vor fi suficient de

puternice pentru a ne face rău (pentru că vor fi) sau dacă vor acţiona întotdeauna în interesele noastre (pentru că nu o vor face), ci dacă, pe termen lung, ne vor ajuta să ne găsim calea – unde vom ajunge pe continuumul Panaceu/Apocalipsă".

Am vorbit despre maşini, dar nu doar despre maşini cu creiere de alamă şi muşchi de fier. Când atomii umani sunt croşetaţi într-un organism în care sunt folosiţi, nu în dreptul lor deplin de fiinţe umane responsabile, ci ca rotiţe, manete şi tije, contează foarte puţin că materia lor primă sunt carnea şi oasele. Ce este utilizat ca element într-o maşină este, de fapt, un element într-o maşină. Indiferent dacă încredinţăm deciziile noastre maşinilor de metal sau maşinilor din came şi oase care sunt birourile, laboratoarele vaste, armatele şi corporaţiile, nu vom primi niciodată răspunsul corect la întrebările noastre decât dacă punem întrebările potrivite… Este foarte târziu şi alegerea între bine şi rău bate la uşă.

Norbert Wiener, The Human Use of Human Beings

Norbert Wiener a fost un profet în ceea ce priveşte recunoaşterea potenţialelor pericole create de maşinile inteligente emergente. Eu consider că a fost şi mai profetic recunoscând că primele inteligenţe artificiale erau deja emergente. A identificat corect corporaţiile şi birourile pe care le numea „maşini în carne şi oase" ca fiind primele maşini inteligente. A anticipat pericolele creării superinteligenţelor artificiale care ar putea avea obiective ce nu sunt neapărat aliniate cu ale noastre.

Ce este clar acum, indiferent dacă era sau nu clar şi pentru Wiener, este că aceste superinteligenţe organizaţionale nu sunt făcute doar din oameni, ci sunt hibrizi între oameni şi tehnologiile informatice care le permit coordonarea. Chiar şi pe vremea lui Wiener, „birourile, laboratoarele vaste, armatele şi corporaţiile" nu puteau funcţiona fără telefoane, telegrafe, radiouri şi tabulatoare. Astăzi nu ar putea funcţiona fără reţele de computere, baze de date şi sisteme de asistenţă în luarea deciziilor. Aceste inteligenţe hibride sunt reţele de oameni augmentate tehnologic. Aceste inteligenţe artificiale au puteri supraomeneşti. Pot şti mai mult decât indivizii în parte, pot simţi mai mult, pot face analize mai complicate şi planuri mai complexe. Pot avea mult mai multe resurse şi mai multă putere decât orice individ în parte.

Deşi nu percepem întotdeauna acest lucru, superinteligenţele hibride cum ar fi statele naţionale şi corporaţiile au propriile lor obiective emergente. Deşi sunt construite deşi pentru oameni, adesea acţionează ca nişte entităţi inteligente independente şi acţiunile lor nu sunt aliniate întotdeauna cu interesele oamenilor care le-au creat. Statul naţional nu este întotdeauna pentru cetăţean şi nici compania pentru acţionar. La fel, organizaţiile nonprofit, ordinele religioase şi

partidele politice funcţionează întotdeauna pentru promovarea principiilor fondatoare. Intuitiv, recunoaştem că acţiunile lor sunt ghidate de obiective interne, motiv pentru care le personificăm, atât legal, cât şi în obiceiurile noastre de gândire. Când vorbim despre „ce vrea China" sau „ce încearcă General Motors să facă", nu vorbim în metafore. Aceste organisme acţionează ca nişte inteligenţe care percep, iau decizii şi acţionează. La fel ca obiectivele oamenilor, obiectivele organizaţiilor sunt complexe şi adesea contradictorii, dar sunt obiective adevărate în sensul că direcţionează acţiunea. Aceste obiective depind oarecum de obiectivele oamenilor din organizaţie, dar nu sunt identice.

Orice american ştie cât de slabă este legătura dintre acţiunile guvernului american şi aspiraţiile diverse şi adesea contradictorii ale cetăţenilor săi. Acelaşi lucru este adevărat şi pentru corporaţii.

Corporaţiile orientate către profit servesc nominal mai multor constituenţi, inclusiv acţionarilor, directorilor de departamente, angajaţilor şi clienţilor. Aceste corporaţii diferă în modul în care jonglează cu loialitatea şi se comportă adesea în moduri care nu servesc niciunuia dintre constituenţii lor. „Neuronii" care transportă gândirea corporatistă nu sunt doar angajaţii umani sau tehnologiile care îi conectează; sunt, de asemenea, codaţi în regulile, structurile de stimulente, cultura şi obiceiurile procedurale ale corporaţiei. Obiectivele corporatiste emergente nu reflectă întotdeauna valorile oamenilor care le implementează. De exemplu, o companie petrolieră condusă deşi în care lucrează oameni cărora le pasă de mediu poate avea structuri de stimulente sau reguli care o vor face să compromită siguranţa mediului de dragul câştigurilor corporative. Intenţiile bune ale componenţilor nu sunt o garanţie pentru comportamentul bun emergent al sistemului.

Guvernele şi corporaţiile, ambele compuse parţial din oameni, sunt motivate natural să pară cel puţin că împărtăşesc obiectivele oamenilor de care depind. Nu ar putea funcţiona fără oameni, deci trebuie să-i ţină cooperanţi. Când astfel de organizaţii par să se comporte altruist, aceasta este adesea o parte importantă din motivaţie. Am complimentat la un moment dat CEO-ul unei corporaţii mari pentru contribuţia pe care a făcut-o compania sa pentru un efort de întrajutorare umanitară. CEO-ul a răspuns fără urmă de ironie: „Da. Am hotărât să facem mai multe lucruri ca acela pentru a face marca noastră mai plăcută". Indivizii care compun o superinteligenţă hibridă pot exercita uneori o influenţă de „umanizare" – de exemplu, un angajat poate încălca regulile companiei pentru a răspunde nevoilor altui om. Angajatul poate acţiona din empatie veritabilă, dar nu trebuie să atribuim o astfel de empatie superinteligenţei în sine. Maşinile hibride au obiective şi cetăţenii/clienţii/angajaţii lor sunt

câteva dintre resursele pe care le folosesc pentru a le atinge.

Suntem aproape de a putea construi superinteligenţe dintr-o tehnologie a informaţiei pură, fără componente umane. La asta se referă oamenii în general când spun „inteligenţă artificială”. Este rezonabil să întrebăm care vor fi atitudinile acestor maşini superinteligente ipotetice faţă de oameni. Îi vor vedea şi ele pe oameni ca pe nişte resurse utile şi importanţa păstrării unor relaţii bune? Vor fi construite astfel încât să aibă obiective aliniate cu ale noastre? O superinteligenţă ar sesiza importanţa acestor întrebări? Care sunt „întrebările potrivite” pe care ar trebui să ni le punem? Cred că una dintre cele mai importante este următoarea: ce relaţie vor avea diversele superinteligenţe una cu cealaltă?

Este interesant să ne gândim cum se descurcă în prezent superinteligenţele hibride cu conflictele dintre ele. Astăzi o mare parte dintre puterile depline stă în braţele statelor naţionale, care pretind autoritate peste un petic de pământ. Indiferent dacă sunt optimizate pentru a acţiona în interesul cetăţenilor lor sau în cel al unui conducător despotic, statele naţionale îşi arogă prioritate asupra dorinţelor şi obiectivelor altor inteligenţe din zona geografică controlată. Pretind monopol asupra folosirii forţei şi recunosc doar alte state naţionale ca egali. Sunt dispuse, dacă este cazul, să solicite sacrificii mari din partea cetăţenilor pentru a-şi asigura autoritatea, chiar până la sacrificarea vieţilor cetăţenilor.

Această împărţire geografică a autorităţii avea sens când majoritatea actorilor erau oameni care-şi petreceau întreaga viaţă într-un singur stat naţional, dar acum, când actorii importanţi includ inteligenţe hibride distribuite geografic, cum ar fi corporaţiile multinaţionale, această logică este mai puţin evidentă. Astăzi trăim într-o perioadă de tranziţie complexă, când superinteligenţele distribuite încă se bazează în mare parte pe statele naţionale pentru a rezolva problemele care iau naştere între ele. Adesea aceste probleme sunt rezolvate diferit în diferite jurisdicţii. Devine din ce în ce mai greu să atribuim persoanele individuale statelor naţionale: călătorii internaţionali care trăiesc şi lucrează în afara ţării lor de origine, refugiaţii şi imigranţii (cu sau fără acte) încă sunt priviţi ca nişte excepţii stingheritoare. Superinteligenţele construite strict din tehnologia informaţiei se vor dovedi şi mai stingheritoare pentru sistemul teritorial de autoritate, pentru că nu va mai exista niciun motiv să fie legate de resursele fizice dintr-o singură ţară – de altfel, de nicio resursă fizică anume. O inteligenţă artificială poate exista bine-mersi „în *clotuty* şi nu într-un anumit loc fizic.

Pot să-mi imaginez cel puţin patru scenarii pentru modul în care superinteligenţele-maşină vor relaţiona cu superinteligenţele hibride.

Într-un scenariu evident, mai multe inteligenţe maşină vor fi

controlate şi aliate în cele din urmă cu statele naţionale individuale. În acest scenariu stat/inteligenţă artificială, ne putem imagina superinteligenţele artificiale americane şi chineze luptându-se pentru resurse în numele statului lor. Într-un anumit sens, aceste inteligenţe artificiale ar fi cetăţeni ai statelor lor naţionale în modul în care multe corporaţii comerciale acţionează adesea drept „cetăţeni corporativi" astăzi. În acest scenariu, statele naţionale gazdă probabil ar conferi superinteligenţelor-maşină resursele de care au nevoie pentru a lucra în avantajul statului. Or, în măsura în care superinteligenţele pot influenţa guvernele lor naţionale, vor face asta pentru a-şi spori propriile puteri, de exemplu colectând o mare parte din resursele statului. Inteligenţele artificiale ale statelor naţionale probabil nu vor vrea ca în jurisdicţia lor să crească inteligenţe artificiale concurente. În acest scenariu, superinteligenţele devin o extensie a statului şi viceversa.

Scenariul stat/inteligenţă artificială pare plauzibil, dar nu este cursul nostru curent. Cele mai puternice inteligenţe artificiale ale noastre şi cele care se îmbunătăţesc cel mai rapid sunt cele controlate de corporaţiile orientate pe profit. Acesta este scenariul corporaţie/ inteligenţă artificială, în care echilibrul puterii între statele naţionale şi corporaţii se inversează. Astăzi cele mai puternice şi inteligente colecţii de maşini sunt probabil deţinute de Google, dar companii precum Amazon, Baidu, Microsoft, Facebook, Apple şi IBM nu sunt, probabil, cu mult în urmă. Aceste companii văd toate un imperativ comercial în a construi inteligenţe artificiale ele însele. Este uşor să ne imaginăm un viitor în care corporaţiile îşi construiesc propriile inteligenţe-maşină, protejate defirewall-*un* şi împiedicând maşinile să profite una de cunoştinţele celeilalte. Aceste maşini vor fi proiectate astfel încât să aibă obiective aliniate cu cele ale corporaţiilor. Dacă această aliniere este eficientă, statele naţionale ar putea continua să rămână în urmă în dezvoltarea propriilor capacităţi de inteligenţă artificială depinzând în schimb de „cetăţenii lor corporatişti" pentru a face acest lucru în locul lor. În măsura în care corporaţiile îşi controlează cu succes obiectivele, vor deveni mai puternice şi mai autonome decât statele naţionale.

Alt scenariu, poate cel de care oamenii se tem cel mai mult, este că inteligenţele artificiale nu vor fi aliniate cu inteligenţele omeneşti sau cu superinteligenţele hibride, ci vor acţiona strict în propriul interes. S-ar putea chiar uni într-o superinteligenţă-maşină unică, de vreme ce nu va exista nicio cerinţă tehnică pentru ca inteligenţele-maşină să-şi păstreze identităţile individuale. Atitudinea unei superinteligenţe artificiale interesate doar de sine faţă de superinteligenţele hibride va fi probabil una competitivă. Oamenii ar putea fi priviţi ca nişte obstacole mărunte, ca furnicile la un picnic,

dar superinteligenţele hibride – cum ar fi corporaţiile, religiile organizate şi statele naţionale – ar putea fi ameninţări existenţiale. La fel ca superinteligenţele hibride, inteligenţele artificiale i-ar putea vedea pe oameni în principiu ca pe nişte instrumente utile pentru atingerea obiectivelor lor, ca pe nişte pioni în competiţia lor cu alte superinteligenţe. Sau am putea fi pur şi simplu irelevanţi. Nu este imposibil ca o inteligenţă maşină să fi luat deja naştere, iar noi să nu o recunoaştem. Poate că nu-şi doreşte să fie observată sau poate fi atât de străină pentru noi încât să nu o putem percepe. Acest lucru face ca scenariul inteligenţei artificiale interesate de sine să fie cel mai greu de imaginat. Cred că versiunile uşor de imaginat, cum ar fi roboţii inteligenţi umanoizi din science-fiction, sunt cel mai puţin probabile. Cele mai complexe maşini ale noastre, de exemplu internetul, au crescut deja dincolo de înţelegerea detaliată a unui singur om, iar comportamentele lor emergente ar putea fi mult în afara sferei noastre de cunoştinţe.

Ultimul scenariu este cel în care inteligenţele-maşină nu vor fi aliate una cu cealaltă, ci vor lucra să atingă obiectivele umanităţii ca întreg. În acest scenariu optimist, inteligenţa artificială ne-ar putea ajuta să restaurăm echilibrul puterii între indivizi şi corporaţii, între cetăţean şi stat. Ne-ar putea ajuta să rezolvăm problemele care au fost create de superinteligenţele hibride care subminează obiectivele oamenilor. În acest scenariu, inteligenţele artificiale ne vor conferi putere, oferindu-ne acces la capacitatea de procesare şi la cunoştinţele disponibile în prezent doar statelor şi corporaţiilor. Prin urmare, ar putea deveni extinderi ale propriilor noastre inteligenţe, pentru atingerea obiectivelor noastre umane. Ar putea face inteligenţele noastre individuale slabe puternice. Această perspectivă este revigorantă şi plauzibilă. Este plauzibilă pentru că avem de ales în ceea ce construim şi avem obiceiul de a folosi tehnologia pentru a extinde şi augmenta capacităţile umane. La fel cum avioanele ne-au dat aripi şi motoarele ne-au dat muşchi să mutăm munţii din loc, la fel reţelele noastre de computere ne pot amplifica şi extinde minţile. Poate că nu ne înţelegem sau controlăm pe deplin destinul, dar avem şansa să îl orientăm în direcţia valorilor noastre. Viitorul nu este ceva ce ni se va întâmpla; este ceva ce vom construi noi.

De ce a văzut Wiener ce nu au văzut alţii

În ingineria electrică, există o divizare care este cunoscută în Germania drept divizarea între tehnica curenţilor puternici şi tehnica curenţilor slabi şi pe care noi o cunoaştem ca distincţia dintre ingineria electrică şi ingineria telecomunicaţiilor. Această divizare separă epoca ce tocmai s-a încheiat şi epoca în care trăim acum.

Norbert Wiener,
Cibernetica sau controlul şi comunicaţiile la animale şi maşini

Cibernetica este studierea modului în care ceva slab controlează ceva puternic. Gândeşte-te la metafora definitorie a domeniului: timonierul care ghidează nava cu ajutorul timonei. Obiectivul timonierului este să controleze direcţia navei, să o ţină pe cursul potrivit. Informaţia, mesajul care este transmis timonierului, vine de la busolă sau de la stele, iar timonierul închide ciclul de *feedback* trimiţând mesajele de ghidare prin forţa fină a mâinii sale către timonă. În această imagine, vedem nava bătută de vânturi şi valuri puternice în lumea reală, controlată de sistemul de comunicare al mesajelor din lumea informaţiei.

Însă distincţia dintre „real" şi „informaţie" este în cea mai mare parte o diferenţă în perspectivă. Semnalele care transportă mesaje, cum ar fi lumina stelelor şi presiunea mâinii pe timonă, există într-o lume de energie şi forţe, la fel şi timonierul. Forţele slabe care controlează cârma sunt la fel de reale şi de fizice precum forţele puternice ce bat în navă. Dacă schimbăm perspectiva noastră cibernetică de la navă la timonier, presiunea cârmei devine o forţă puternică a muşchilor controlată de semnalele slabe ale minţii timonierului. Mesajele din mintea timonierului sunt amplificate într-o forţă fizică suficient de puternică încât să ghideze nava. Sau, în schimb, putem mări distanţa focală a imaginii şi putem vedea perspectiva de ansamblu a ciberneticii. Putem vedea nava ca parte dintr-o reţea comercială vastă, parte dintr-un ciclu cu *feedback* ce reglementează preţul mărfurilor prin fluxul de bunuri. În această perspectivă, mica navă este doar un mesager. Astfel că distincţia dintre lumea fizică şi lumea informaţiilor este modul în care descriem relaţia dintre slab şi puternic.

Wiener a ales să vadă lumea din punctul de vedere şi la scara omului individual. Ca cibernetician, a preluat perspectiva protagonistului slab integrat într-un sistem puternic, încercând să profite la maximum de puterile sale limitate. A încorporat această perspectivă în propria sa definiţie a informaţiei. „Informaţia", spunea el, „este numele conţinutului a ceea ce este schimbat cu lumea exterioară pe măsură ce ne potrivim ei şi facem această potrivire resimţită şi asupra ei". În cuvintele sale, informaţia este ceea ce folosim pentru a „trăi efectiv în acel mediu"[56]. Pentru Wiener, informaţia este modalitatea în care slabul face faţă puternicului. Acest punct de vedere se reflectă şi în definiţia lui Gregory Bateson a informaţiei ca „diferenţă ce face diferenţa", prin care se referea la micile diferenţe care determină o diferenţă mare.

Obiectivul ciberneticii era să creeze un model mic al sistemului folosind „curenţii slabi" pentru a amplifica şi controla „curenţii

puternici" din lumea reală. Observaţia centrală era că o problemă de control putea fi soluţionată construind un sistem analog în spaţiul de informaţie al mesajelor şi apoi amplificând soluţiile în lumea mai largă a realităţii. Inerent mişcării unui sistem de control este conceptul de amplificare ce face din mic mare şi din slab puternic. Amplificarea permite diferenţa care face diferenţa în a face o diferenţă.

În acest mod de a privi lumea, un sistem de control trebuie să fie la fel de complex ca sistemul pe care îl controlează. Cibernetician ul W. Ross Ashby a dovedit că acest lucru este adevărat într-un sens matematic şi precis, în ceea ce se numeşte acum

Legea lui Ashby privind diversitatea necesară sau, uneori, prima lege a ciberneticii. Legea ne spune că, pentru a controla complet un sistem, controlorul trebuie să fie la fel de complex precum controlatul. Cibernet icienii au tendinţa să vadă sistemele de control ca un fel de analog pentru sistemele pe care le controlează, ca homunculusul – mica persoană ipotetică din creier care controlează persoana fizică.

Această noţiune de structură analogică este uneori confundată cu noţiunea codării analogice a mesajelor, dar cele două sunt diferite din punct de vedere logic. Norbert Wiener a fost foarte impresionat de analizorul diferenţial digital al lui Vannevar Bush, care putea fi reconfigurat pentru a se potrivi structurii oricărei probleme îi era dată spre rezolvare, dar folosea o codare digitală a semnalului. Semnalele puteau fi simplificate pentru a reprezenta deschis distincţiile relevante, permiţându-le să fie comunicate şi stocate cu o precizie mai mare. În semnalele digitale, trebuia păstrată doar diferenţa semnalelor care constituiau diferenţa. Această diferenţă şi codare de semnal este utilizată în prezent pentru a diferenţia între „analog" şi „digital". Codarea digitală a semnalului era complet compatibilă cu gândirea cibernetică – de fapt, aceasta o făcea posibilă. Ce era restrictiv pentru cibernetică era presupunerea unei analogii de structură între controlor şi controlat. Până în 1930, Kurt Godel, Alonzo Church şi Alan Turing descriseseră cu toţii sisteme universale de calcul, în care tehnica de calcul nu necesita nicio analogie structurală pentru funcţiile calculate. Aceste computere universale puteau, de asemenea, să calculeze funcţiile de control.

Analogia de structură dintre controlor şi controlat a fost centrală pentru perspectiva ciberneticii. La fel cum codarea digitală comprimă spaţiul mesajelor posibile într-o versiune simplificată ce reprezintă doar diferenţa ce face o diferenţă, la fel şi sistemul de control comprimă spaţiul stării sistemului controlat într-un model simplificat ce reflectă doar obiectivele controlorului. Legea lui Ashby nu implică faptul că orice controlor trebuie să modeleze fiecare stare a sistemului, ci doar acele stări care contează pentru avansarea obiectivelor controlorului. Prin urmare, în cibernetică, obiectivul controlorului

devine perspectiva din care este văzută lumea.

Norbert Wiener a adoptat perspectiva omului individual relaţionând cu organizaţii vaste şi încercând să „trăiască efectiv în acel mediu". A preluat perspectiva celui slab încercând să-l influenţeze pe cel puternic. Poate că acesta este motivul pentru care a putut observa obiectivele emergente ale „maşinilor în carne şi oase" şi anticipa câteva dintre provocările umane ridicate de aceste inteligenţe noi – inteligenţe maşinice hibride cu obiective proprii.

Capitolul 18
Vor deveni computerele stăpânii noştri

Venki Ramakrishnan

Venki Ramakrishnan este cercetător la Laboratorul de Biologie Moleculară al Consiliului de Cercetare Medicală de la Universitatea Cambridge, laureat al Premiului Nobel în chimie (2009), preşedintele actual al Societăţii Regale şi autorul cărţii Gene Machine: The Race to Discover the Secrets of the Ribosome.

Venki Ramakrishnan este un biolog laureat al Premiului Nobel ale cărui contribuţii ştiinţifice includ activitatea sa în domeniul structurii atomice a ribozomilor – care este, efectiv, o uriaşă maşină moleculară care ne citeşte genele şi face proteine. Activitatea sa ar fi fost imposibilă fără computere puternice. Internetul i-a uşurat mult munca şi, observă el, a acţionat ca un element omogenizator internaţional: „Când am crescut în India, dacă voiai să faci rost de o carte, apărea cu şase luni sau chiar un an după ce vedea lumina tiparului în Occident… Revistele ajungeau prin poşta obişnuită după câteva luni. Nu a trebuit să fac faţă acestei probleme, pentru că am plecat din India la vârsta de nouăsprezece ani, dar cunosc cercetători indieni care au trebuit să-i facă faţă. Astăzi, au acces la informaţie printr-un simplu clic. Mai important, au acces la prelegeri. Îl pot asculta pe Richard Feynman. Acesta ar fi fost un vis al meu din copilărie. Îl pot urmări pur şi simplu pe Richard Feynman pe internet. Aceasta este o nivelare foarte puternică în domeniu". Cu toate acestea… „alături de beneficiile internetului, acum a apărut un volum uriaş de zgomot. Apar tot soiul de oameni care perorează în jargon pseudoştiinţific, promovându-şi ideile ca şi cum ar fi ştiinţifice".

În calitate de preşedinte al Societăţii Regale, Venki se îngrijorează şi el despre problema mai largă a încrederii: încrederea publică în studiile ştiinţifice bazate pe dovezi, dar şi încrederea în rândurile oamenilor de ştiinţă, consolidată de

verificările riguroase ale concluziilor unui coleg de breaslă – încredere care este în pericol de erodare din cauza caracterului de „cutie neagră" al computerelor cu învăţare profundă. „Această eroziune va apărea din ce în ce mai mult, pe măsură ce seturile de date se măresc, pe măsură ce vom avea studii generale despre genom, studii pe populaţie şi tot soiul de alte studii", spune el. „Cum vom face faţă noi, comunitatea ştiinţifică, acestui lucru şi cum vom comunica publicului o idee despre esenţa ştiinţei, ce este fiabil în ştiinţă, ce este nesigur în ştiinţă şi ce este pur şi simplu greşit în ştiinţă?"

Un fost coleg al meu, Gérard Bricogne, obişnuia să glumească spunând că inteligenţa bazată pe carbon este doar un catalizator pentru evoluţia inteligenţei bazate pe siliciu. Mult timp, atât filmele de la Hollywood, cât şi profeţii ştiinţifici au preconizat capitularea noastră în faţa stăpânilor noştri, computere. Cu toţii aşteptăm singularitatea, care pare întotdeauna să fie după colţ.

Într-un anumit sens, computerele au preluat deja controlul, facilitând practic fiecare aspect al vieţilor noastre – de la banking, călătorii şi utilităţi până la cele mai intime comunicări personale. Pot să-l văd şi să vorbesc *gratuit* cu nepotul meu din New York. Îmi amintesc când am văzut prima dată filmul din 1968 intitulat *2001: Odiseea spaţială*, publicul a râs când a auzit preţul ridicol de ieftin de 1, 7 dolari pentru un apel telefonic cu imagini venite din spaţiu, în acel moment un apel la distanţă costând 3 dolari pe minut în Statele Unite.

Totuşi, comoditatea şi puterea computerelor este şi ceva asemănător unui târg faustic, deoarece vine la pachet cu pierderea controlului. Computerele ne împiedică să facem lucrurile pe care vrem să le facem. Încearcă să te urci într-un avion dacă ajungi la aeroport şi sistemele de computere nu funcţionează, aşa cum i s-a întâmplat cu puţin timp în urmă companiei British Airways la aeroportul Heathrow. Avioanele, piloţii şi pasagerii erau toţi acolo, chiar şi controlul traficului aerian funcţiona. Dar niciun avion al companiei aeriene nu putea decola. Computerele ne fac şi să facem lucruri pe care *nu vrem* să le facem – generând liste de e-mailuri şi tipărind etichete pentru a ne trimite milioane de e-mailuri nedorite pe care noi, oamenii, trebuie să le sortăm, expediem şi ştergem.

Dar nu aţi văzut nimic încă. În trecut, am programat computerele folosind algoritmi pe care i-am înţeles cel puţin în principiu. Şi, când maşinile au obţinut rezultate uimitoare, cum ar fi să-l învingă pe campionul mondial la şah, Garry Kasparov, am putut spune că programele victorioase au fost proiectate cu algoritmi bazaţi pe propria noastră înţelegere – folosind, în acest caz, experienţa şi sfaturile celor mai buni jucători de şah. Maşinile erau pur şi simplu

mai rapide la efectuarea calculelor brute, aveau volume prodigioase de memorie şi nu erau pasibile de eroare. Un articol a descris victoria lui Deep Blue nu ca pe cea a unui computer, care era doar o maşină proastă, ci ca victoria a sute de programatori asupra lui Kasparov, un singur om.

Acest mod de programare se modifică dramatic. După o discontinuitate mare, puterea învăţării automate a prins aripi. O mare parte din schimbare a venit atunci când programatorii, în loc să încerce să anticipeze şi să codeze toate contingenţele posibile, au permis computerelor să se instruiască singure în ce priveşte datele, folosind reţele neuronale profunde bazate pe modele pentru cum învaţă creierele noastre. Folosesc metode probabilistice pentru „a învăţa" din volume mari de date. Computerele pot recunoaşte tipare şi pot trage concluzii fără ajutorul oamenilor. O metodă foarte puternică pare să fie învăţarea cu întărire, prin care computerul învaţă, fără intrări anterioare, ce variabile sunt importante şi câtă pondere să le acorde pentru a atinge un anumit obiectiv. Această metodă imită într-o oarecare măsură modul în care învăţăm când suntem copii. Rezultatele acestor abordări noi sunt uimitoare.

Un astfel de program de învăţare profundă a fost folosit pentru a învăţa un computer să joace Go, un joc despre care doar până acum câţiva ani se credea că nu poate fi jucat de o inteligenţă artificială pentru că era atât de greu să estimezi cât de bine stai. Se pare că cei mai buni jucători de Go se bazau foarte mult pe intuiţie şi pe estimarea poziţiei, deci se considera că priceperea necesita un anumit tip de inteligenţă umană. Dar programul Alpha Go produs de Deep Mind, după ce a fost instruit din mii de jocuri Go de nivel superior jucate de oameni şi apoi de milioane de jocuri pe care le-a jucat cu el însuşi, a putut foarte repede să-i bată pe cei mai buni jucători din lume. Şi mai uimitor este faptul că programul corespunzător Alpha Go Zero, care a învăţat totul de la zero jucând singur, a fost mai puternic decât versiunea instruită iniţial pe jocurile cu oameni! Este ca şi cum oamenii au împiedicat computerul să ajungă la potenţialul său maxim. Aceeaşi metodă a fost recent generalizată: începând de la zero, în numai douăzeci şi patru de ore, un program echivalent de şah de la Alpha Zero a reuşit să învingă cele mai bune programe de şah „convenţionale" de astăzi, care, la rândul lor, au învins cei mai buni oameni.

Progresul nu a fost restricţionat la jocuri. Computerele sunt semnificativ mai bune la recunoaşterea imaginilor şi la sintetizarea vocii decât erau înainte. Pot detecta tumorile din radiografii mai devreme decât majoritatea oamenilor. Diagnosticele medicale şi medicamentele personalizate se vor îmbunătăţi substanţial. Transportul cu maşinile autonome ne va menţine pe toţi în siguranţă,

în medie. Probabil nepotul meu nu va trebui să-şi ia niciodată permisul de conducere, pentru că a conduce o maşină va fi cum este astăzi călăritul – un hobby pentru câţiva oameni. Activităţile periculoase cum ar fi minatul şi activităţile repetitive şi plictisitoare vor fi făcute de computere. Statul va oferi servicii publice mai precise, mai personalizate şi mai eficiente. Inteligenţa artificială poate revoluţiona educaţia analizând toate nevoile individuale ale elevilor şi permiţând predarea personalizată, astfel încât toţi elevii să poată avansa la o rată optimă.

Alături de aceste beneficii uriaşe, bineînţeles, vom avea şi riscuri alarmante. Cu volumul vast de date personale, computerele vor putea afla mai multe despre noi decât ştim noi înşine despre noi; cine deţine datele despre noi va deveni o întrebare foarte importantă. În plus, deciziile bazate pe date vor reflecta fără îndoială predilecţiile sociale: chiar şi un aşa-zis sistem inteligent neutru conceput pentru prezicerea riscurilor de credit, să spunem, poate trage concluzia că simpla calitate de membru al unui grup minoritar te predispune la incapacitatea de rambursare a împrumutului. Deşi acesta este un exemplu evident pe care-l putem corecta, pericolul real este că nu suntem întotdeauna conştienţi de predilecţiile din date şi le putem perpetua pur şi simplu.

De asemenea, învăţarea automată poate perpetua propriile noastre prejudecăţi. Când Netflix sau Amazon încearcă să-ţi spună ce ţi-ar plăcea să vizionezi sau să cumperi, aceasta este o aplicaţie a învăţării automate. În prezent, astfel de sugestii sunt uneori ridicole, dar cu timpul şi cu mai multe date, vor fi din ce în ce mai precise, fortificându-ne prejudecăţile, ce ne place şi ce nu ne place. Vom rata oare ocazia întâlnirii aleatorii care ne-ar putea convinge să ne schimbăm părerea prin expunere la idei noi şi contradictorii? *Social-mediay* dată fiind influenţa sa asupra alegerilor, este un exemplu destul de clar pentru modul în care poate fi accentuată falia dintre oamenii care se află în părţi diferite ale spectrului politic.

Probabil am atins deja etapa în care majoritatea guvernelor nu mai au puterea să reziste la influenţa combinată a câtorva companii multinaţionale puternice care ne controlează pe noi şi viitorul nostru digital. Lupta dintre companiile dominante de astăzi este, de fapt, o luptă pentru controlarea datelor. Acestea îşi vor folosi influenţa lor enormă pentru a împiedica reglementarea datelor, pentru că sunt interesate de controlul absolut asupra lor. În plus, posedă resursele financiare pentru a-i angaja pe cei mai talentaţi oameni din domeniu, sporindu-şi astfel şi mai mult puterea. Ne-am cedat datele personale în schimbul câtorva servicii gratuite precum Gmail şi Facebook, dar, după cum evidenţia autorul şi jurnalistul John Lanchester în *London Review of Books*, dacă este gratuit, atunci tu eşti produsul. Clienţii lor

adevăraţi sunt cei care-i plătesc pentru a primi acces la cunoştinţele acumulate despre noi, pentru a ne putea convinge să le cumpărăm produsele sau pentru a ne influenţa în altă privinţă. O metodă de a evita monopolul asupra datelor este să împărţim titlul de proprietate asupra datelor unor firme care le utilizează. Persoanele individuale ar deţine şi controla astfel datele lor cu caracter personal (un model care ar încuraja competiţia, devreme ce oamenii ar fi liberi să-şi mute datele la compania care oferă servicii mai bune). În cele din urmă, folosirea abuzivă a datelor nu este limitată la corporaţii: în regimurile totalitare sau chiar şi în cele democrate cu numele, guvernele ştiu lucruri despre cetăţenii lor pe care nici Orwell nu şi le-ar fi putut imagina. Modul în care utilizează aceste informaţii nu este întotdeauna transparent şi nu poate fi contracarat mereu.

Ideea de inteligenţă artificială în scopuri militare este înspăimântătoare. Ne putem imagina sisteme inteligente care sunt proiectate să acţioneze autonom în baza datelor obţinute în timp real, care pot acţiona mult mai repede decât inamicul, iniţiind războaie catastrofale. Astfel de războaie nu sunt neapărat convenţionale sau nucleare. Dată fiind importanţa reţelelor de computere în societatea modernă, este mult mai probabil ca războaiele inteligenţelor artificiale să fie purtate în spaţiul cibernetic. Consecinţele pot fi la fel de grave.

În ciuda acestei pierderi a controlului, continuăm să mărşăluim inexorabil într-o lume în care inteligenţa artificială va fi peste tot. Indivizii nu vor putea rezista comodităţii şi puterii, iar corporaţiile şi guvernele nu vor putea rezista avantajelor sale competitive. Dar iau naştere întrebări importante legate de viitorul muncii. Computerele sunt responsabile de pierderi considerabile în rândul meseriilor de muncă manuală în ultimele decenii, dar, până de curând, meseriile de birou – meserii pe care „le pot face doar oamenii” – erau considerate sigure. Brusc, acest lucru nu mai este adevărat. Contabili, mulţi profesionişti din domeniul dreptului sau medicinei, analişti financiari şi brokeri, agenţi de turism – de fapt, mulţi angajaţi la birou – vor dispărea ca urmare a programelor sofisticate de învăţare automată. Vom face faţă unui viitor în care fabricile vor scoate pe bandă rulantă produse cu foarte puţini angajaţi, iar transportul bunurilor este în mare parte automatizat, la fel ca multe alte servicii. Ce mai rămâne de făcut pentru oameni?

În 1930 – cu mult timp înainte de apariţia computerelor, cu atât mai mult a inteligenţei artificiale –, John Maynard Keynes scria, într-un eseu numit „Posibilităţile economice ale nepoţilor noştri”, că, din cauza îmbunătăţirilor productivităţii, societatea poate produce toate lucrurile de care are nevoie cu o săptămână de lucru de cincisprezece ore. De asemenea, a preconizat, odată cu creşterea timpului liber

petrecut în mod creativ, sfârşitul banilor şi avuţiei ca scop în viaţă:

> *Vom putea să ne permitem să îndrăznim să evaluăm motivul banilor la valoarea sa adevărată. Dragostea faţă de bani ca posesie – diferită de dragostea faţă de bani ca mijloc de obţinere a plăcerilor şi realităţilor vieţii – va fi recunoscută pentru ceea ce este, o morbiditate oarecum dezgustătoare, una dintre acele înclinaţii semicriminale, semipatologice pe care le predăm ridicând din umeri specialiştilor în boli mintale.*

Din păcate, predicţia lui Keynes nu s-a adeverit. Deşi productivitatea a crescut, într-adevăr, sistemul – posibil inerent într-o economie de piaţă – nu a avut ca efect mai puţine ore de lucru. În schimb, s-a întâmplat ceea ce antropologul şi anarhistul David Graeber a descris ca fiind creşterea „meseriilor de rahat”[57]. În timp ce meseriile care produc elemente esenţiale precum alimente, adăpost şi bunuri au fost în mare parte automatizate, am văzut o extindere enormă a unor sectoare precum legislaţia corporatistă, administrarea academică şi a sănătăţii (în opoziţie cu predatul efectiv, cercetarea şi practica medicinei), „resursele umane” şi relaţiile publice, fără a mai menţiona industriile noi precum serviciile financiare şi telemarketingul, şi industriile auxiliare în aşa numita economie a gigurilor, care serveşte persoanelor care sunt prea ocupate făcând toată munca suplimentară.

Cum se vor descurca societăţile cu tehnologia care distruge rapid întregi profesii şi lasă un număr mare de oameni fără loc de muncă? Unii argumentează că această grijă se bazează pe o premisă falsă, pentru că iau naştere meserii care nu existau înainte, dar, subliniază Graeber, aceste meserii noi nu vor fi neapărat satisfăcătoare şi nu ne vor face să ne simţim împliniţi. În timpul primei revoluţii industriale, a durat aproape un secol înainte ca mulţi oameni să aibă o situaţie mai bună. Acea revoluţie a fost posibilă doar pentru că guvernele vremii favorizau acerb drepturile de proprietate asupra muncii şi cei mai mulţi oameni (şi toate femeile) nu aveau dreptul de vot. În societăţile democratice de astăzi, nu este clar dacă populaţia va tolera o răsturnare atât de spectaculoasă a societăţii în baza promisiunii că, „în cele din urmă”, lucrurile se vor îmbunătăţi.

Chiar şi această viziune edulcorată va depinde de reformarea radicală a educaţiei şi a procesului de învăţare de-a lungul vieţii. Revoluţia industrială a determinat schimbări sociale enorme de acest tip, inclusiv o alunecare către educaţia universală. Dar nu se va întâmpla decât dacă o facem noi să se întâmple: esenţialmente, este vorba despre putere, mijlocire şi control. Ce urmează pentru, să spunem, şoferul de taxi sau de camion, în vârstă de patruzeci şi şase de ani, în era maşinilor autonome?

O idee care a fost vehiculată este cea a venitului de bază universal, care va permite cetăţenilor să-şi urmărească interesele, să se respectalizeze pentru alte ocupaţii şi, în general, să fie liberi să trăiască o viaţă decentă. Însă economiile de piaţă, care au la bază o cerere din ce în ce mai mare din partea consumatorului mai mult decât orice altceva, probabil nu vor tolera această inovaţie. De asemenea, există senzaţia că a găsi sens în munca prestată este un element esenţial pentru demnitatea umană şi sentimentul de împlinire. Deci altă posibilitate ar fi ca bogăţia enormă generată de productivitatea crescută datorită automatizării să fie redistribuită pentru meseriile ce necesită muncă şi creativitate umane, în domenii precum arta, muzica, asistenţa socială şi alte activităţi ce merită efortul. În cele din urmă, ce meserii vor fi satisfăcătoare sau productive şi care vor fi „de rahat” ţine de părerea fiecăruia şi pot varia de la societate la societate, precum şi în funcţie de timp.

Până acum m-am concentrat pe consecinţele practice ale inteligenţei artificiale. Ca om de ştiinţă, ce mă deranjează pe mine este posibilitatea de a pierde înţelegerea. Acum acumulăm date la o rată incredibilă. Doar în laboratorul meu, un experiment generează mai mult de un terabyte de date pe zi. Aceste date sunt transformate în mesaje, analizate şi reduse până când ajungem la un rezultat interpretabil. Dar în toată această analiză de date, credem că ştim ce se întâmplă. Ştim ce fac programele, pentru că noi suntem cei care am proiectat algoritmii de la zero. Prin urmare, când computerele noastre generează un rezultat, simţim că îl înţelegem intelectual.

Noile programe de învăţare automată sunt diferite. Recunoscând tiparele prin intermediul reţelelor neuronale profunde, generează concluzii şi nu ştim exact cum. Când descoperă relaţii, noi nu le înţelegem în acelaşi fel ca şi cum am fi dedus acele relaţii noi înşine, utilizând un cadru teoretic subiacent. Pe măsură ce seturile de date devin mai mari, nu le vom putea analiza noi înşine, chiar şi cu ajutorul computerelor; mai degrabă ne vom bizui în întregime pe computere să facă analiza pentru noi. Aşadar, dacă cineva ne întreabă cum ştim ceva, vom spune pur şi simplu că maşina a analizat datele şi a produs concluzia.

Într-o bună zi, un computer poate prezenta un rezultat cu totul nou – de exemplu, o teoremă matematică a cărei demonstraţie sau chiar a cărei ipoteză nu poate fi înţeleasă de niciun om. Acesta este un mod filosofic diferit de modul în care am practicat ştiinţa până acum. Sau, cel puţin, de modul în care am crezut că am practicat-o. Unii ar putea argumenta că nu ştim nici cum ajung creierele noastre la anumite concluzii şi că aceste metode sunt o cale de a imita învăţarea realizată de creierul uman. Cu toate acestea, mie mi se pare

înspăimântătoare această posibilă pierdere a înţelegerii.

În ciuda avansurilor remarcabile din domeniul tehnicii de calcul, febra inteligenţei artificiale generale – o maşină generală de inteligenţă care va gândi ca un om şi, probabil, va dezvolta conştiinţă – mie îmi miroase a science-fiction, în mare parte pentru că nu înţelegem creierul la acel nivel de detaliu. Nu doar că nu înţelegem ce este conştiinţa, nu înţelegem nici măcar o problemă relativ simplă, de exemplu, cum reuşim să ţinem minte un număr de telefon. În această întrebare simplă, trebuie să luăm în considerare o mulţime de factori. Cum ştim că este un număr? Cum îl asociem unei persoane, unui nume, unei feţe şi altor caracteristici? Chiar şi astfel de întrebări aparent triviale implică totul de la cogniţia de nivel superior şi memorie până la modul în care o celulă stochează informaţiile şi cum interacţionează neuronii.

În plus, aceasta este doar o sarcină printre multe altele pe care creierul le face fără efort. În timp ce maşinile vor face, fără îndoială, lucruri din ce în ce mai minunate, este puţin probabil să înlocuiască gândirea, creativitatea şi viziunea umane. Eric Schmidt, fostul preşedinte al companiei Google, a spus într-un interviu recent la Muzeul de Ştiinţe din Londra că până şi simpla desemnare a unui robot să cureţe masa, să spele vasele şi să le aşeze va fi o provocare uriaşă. Calculele implicate de a programa toate mişcările pe care trebuie să le facă corpul pentru a arunca o minge cu precizie sau de a face slalom pe schiuri sunt vaste. Creierul poate face toate acestea, făcând în acelaşi timp şi calcule matematice sau muzică şi *inventând* jocuri precum şah şi Go, nu doar să le joace. Avem tendinţa să subestimăm complexitatea şi creativitatea creierului uman şi cât de uimitor de generalist este.

Dacă inteligenţa artificială urmează să devină mai umană în abilităţile sale, comunităţile de învăţare automată şi de neuroştiinţe trebuie să interacţioneze îndeaproape, ceva ce se întâmplă deja. Câteva dintre numele cele mai mari din prezent pentru inteligenţa artificială – cum ar fi Geoffrey Hinton, Zoubin Ghahramani şi Demis Hassabis – au experienţă în neuroştiinţa cognitivă şi succesul lor s-a datorat cel puţin parţial încercărilor de a modela comportamente asemănătoare creierului în algoritmii lor. În acelaşi timp, a înflorit şi neurobiologia. Au fost dezvoltate tot soiul de instrumente pentru a vedea ce neuroni se activează şi pentru a-i manipula genetic şi a vedea ce se întâmplă în timp real cu intrările. Mai multe ţări au lansat iniţiative de neuroştiinţă cu ţeluri foarte înalte pentru a vedea dacă putem să dezvăluim misterele creierului. Progresele în domeniul inteligenţei artificiale şi neuroştiinţei par să meargă mână în mână; fiecare domeniu îl poate avansa pe celălalt.

Mulţi cercetători evoluţionişti şi filosofi precum Daniel Dennett au

scos în evidenţă faptul că creierul uman este rezultatul a miliarde de ani de evoluţie[58]. Inteligenţa umană nu este caracteristica specială pe care o credem noi a fi, ci doar un alt mecanism de supravieţuire, nu foarte diferit de sistemele noastre digestive sau imunitare, şi acestea din urmă fiind ambele incredibil de complexe. Inteligenţa a evoluat pentru că ne-a permis să înţelegem lumea din jurul nostru, să planificăm pentru viitor şi să facem faţă astfel unei game variate de evenimente pentru a supravieţui. Însă, cum scria Descartes, noi, oamenii, ne definim existenţa prin abilitatea noastră de a gândi. Deci nu este surprinzător că, în mod antropomorfic, temerile noastre ce vizează inteligenţa artificială reflectă această credinţă că inteligenţa noastră este cea care ne face speciali.

Dar, dacă facem un pas înapoi şi privim viaţa de pe Pământ, observăm că suntem departe de a fi cea mai rezistentă specie. Dacă vom fi cuceriţi la un moment dat, vom fi cuceriţi de cele mai vechi forme de viaţă de pe Pământ, cum ar fi bacteriile, care pot trăi oriunde, din Antarctica până la vulcanii submarini de pe platforma oceanică, zone mai fierbinţi decât apa clocotită, sau în medii acide care ne-ar topi pe mine şi pe tine. Aşadar, când oamenii întreabă unde ne îndreptăm, trebuie să punem întrebarea într-un context mai larg. Nu ştiu ce fel de viitor va inaugura inteligenţa artificială: dacă îi va face pe oameni sclavii ei sau nişte fiinţe depăşite sau dacă va fi o extindere utilă şi binevenită a capacităţilor noastre, care ne va îmbogăţi vieţile. Dar sunt aproape sigur că computerele nu vor fi niciodată stăpânii bacteriilor.

Capitolul 19
Strategia umană

Alex „Sandy” Pentland

Alex „Sandy” Pentland este profesor Toshiba şi profesor de ştiinţe şi arte media la MIT; director al laboratoarelor de Dinamică Umană şi Ştiinţe ale Conectării şi al programului de Antreprenoriat Media Lab şi autorul cărţii Social Physics.

Alex „Sandy” Pentland, un exponent al ceea ce el a denumit „fizica socială”, este interesat de construirea unor ecologii puternice între om şi inteligenţa artificială. Se preocupă în acelaşi timp de potenţialele pericole ale sistemelor de luare a deciziilor în care datele preiau controlul efectiv asupra creativităţii umane şi creativitatea umană este surghiunită în fundal.

Apariţia Big Data, consideră el, ne-a conferit ocazia de a ne reinventa civilizaţia: „Acum putem începe să privim cu adevărat

detaliile interacţiunii sociale şi ce efecte au acestea, nemaifiind limitaţi la valori medii, cum ar fi indicii de piaţă sau rezultatele alegerilor. Aceasta este o modificare uluitoare. Abilitatea de a vedea detaliile pieţei, ale revoluţiilor politice şi de a le putea preconiza şi controla este cu siguranţă un caz de foc prometeic – poate fi utilizat pentru bine sau pentru rău. Big Data ne-a dus în nişte vremuri interesante".

La întâlnirea noastră de grup din Washington, Connecticut, a mărturisit că, citind despre conceptul de feedback al lui Norbert Wiener, „am simţit că-mi citesc propriile gânduri".

„După Wiener, oamenii au descoperit ori s-au concentrat pe faptul că există sisteme cu adevărat haotice, care pur şi simplu nu sunt predictibile", spunea el, „dar, dacă analizăm sistemele umane socioeconomice, există un procentaj mare de variaţie care poate fi contabilizat şi preconizat… Astăzi avem date de la tot felul de dispozitive digitale şi de la toate tranzacţiile noastre. Faptul că totul este câmp de date înseamnă că poţi măsura lucrurile în timp real în majoritatea aspectelor vieţii umane şi, din ce în ce mai mult, în toate aspectele vieţii omeneşti. Faptul că avem computere şi tehnici de învăţare automată interesante înseamnă că putem construi modele predictive pentru sistemele umane în moduri în care nu se putea face acest lucru până acum".

În ultima jumătate de secol, ideea inteligenţei artificiale şi a roboţilor inteligenţi a dominat gândirea cu privire la relaţia dintre oameni şi computere. Pe de o parte, acest lucru se datorează faptului că este uşor să spui poveşti despre inteligenţa artificială şi roboţi, iar pe de alta, datorită succeselor ei timpurii (de exemplu, demonstratorii de teoreme ce au reprodus majoritatea *principiilor matematice* ale lui Whitehead şi Russell) şi finanţării militare masive. Viziunea mai veche şi mai largă asupra ciberneticii, care considera artificialul ca făcând parte dintre sistemele mai mari de *feedback* şi influenţă reciprocă, s-a estompat din conştiinţa publică.

Însă în anii care au urmat viziunea cibernetică a crescut încet-încet şi a preluat tacit controlul până la punctul în care este „în aer". Cercetarea de ultimă generaţie în majoritatea disciplinelor inginereşti este încadrată acum în sisteme de *feedback* dinamice şi acţionate de fluxuri de energie. Chiar şi inteligenţa artificială este reconstituită sub formă de sisteme „de consiliere" om/maşină şi armata începe finanţarea la scară largă în acest domeniu – ceva ce probabil ar trebui să ne îngrijoreze mai mult decât dronele şi roboţii umanoizi independenţi.

Dar, pe măsură ce ştiinţa şi ingineria au adoptat o postură mai apropiată de cibernetică, a devenit clar că până şi viziunea ciberneticii este mult prea mică. Eram iniţial centraţi pe integrarea actorului

individual, şi nu pe proprietăţile emergente ale unei reţele de actori. Nu este deloc surprinzător, deoarece matematica reţelelor nu a existat până de curând, aşa că o ştiinţă cantitativă a comportamentului reţelelor era imposibilă. Acum ştim că studierea cazurilor individuale nu produce înţelegerea sistemului decât în anumite cazuri simple. Progresele recente în acest domeniu au fost prefigurate de înţelegerea faptului că „haosul" şi, mai târziu, „complexitatea" reprezentau comportamentul tipic al sistemelor, dar acum putem merge mult dincolo de aceste înţelegeri statistice.

Începem să putem analiza, preconiza şi chiar proiecta comportamentele emergente ale reţelelor eterogene complexe. Perspectiva cibernetică a agentului individual conectat poate fi acum extinsă pentru a acoperi sisteme complexe de persoane şi maşini conectate, iar observaţiile pe care le facem în această perspectivă mai largă sunt fundamental diferite de cele făcute din perspectiva cibernetică. Să ne gândim la reţea este analog cu a ne gândi la un întreg ecosistem. Cum ghidezi ecosistemele să crească în direcţia potrivită? Ce înseamnă „direcţia potrivită"? Astfel de întrebări sunt dincolo de graniţa gândirii cibernetice tradiţionale.

Poate că cea mai izbitoare realizare este că oamenii au început deja să utilizeze inteligenţa artificială şi învăţarea automată pentru a ghida întregi ecosisteme, printre care se numără şi multe ecosisteme cu oameni, creând astfel ecologii de oameni şi inteligenţe artificiale la un loc. Acum că totul devine „datificat", putem să măsurăm cele mai multe aspecte ale vieţii omeneşti şi, din ce în ce mai mult, aspecte ale întregii vieţi. Acest lucru, alături de tehnicile de învăţare automată noi şi puternice, înseamnă că putem construi modele ale acestor ecologii altfel decât puteam până acum. Exemplele bine-cunoscute sunt modelele de prognoză meteo şi de preconizare a traficului, care sunt extinse pentru a preconiza clima globală şi pentru a planifica modernizarea şi creşterea oraşelor. Ingineria ajutată de inteligenţa artificială a ecologiilor este deja aici.

Dezvoltarea ecosistemelor om-inteligenţă artificială este, probabil, inevitabilă pentru o specie socială cum suntem noi. Am devenit sociali devreme în evoluţia noastră, cu milioane de ani în urmă. Am început să facem schimb de informaţii pentru a supravieţui, pentru a ne spori adaptarea la mediu. Am dezvoltat scrisul pentru a împărtăşi idei abstracte şi complexe, şi, mai recent, am dezvoltat computerele pentru a ne spori capacităţile de comunicare. Acum dezvoltăm inteligenţa artificială şi modele de învăţare automată pentru ecosisteme, împărtăşind predicţiile acestor modele pentru a ne configura împreună viitorul lumii în care trăim prin legi şi acorduri internaţionale noi.

Trăim un moment istoric fără precedent, în care disponibilitatea unor volume uriaşe de date despre comportamentul uman şi

progresele învăţării automate ne permit să abordăm probleme sociale complexe prin algoritmizarea procesului de luare a deciziilor. Oportunităţile unei astfel de ecologii om-inteligenţă artificială de a avea un impact social pozitiv prin decizii mai echitabile şi mai transparente sunt evidente. Dar există şi riscurile unei „tiranii a algoritmilor", în care experţi în date nealeşi conduc lumea. Alegerile pe care le facem acum sunt, probabil, şi mai importante decât cele pe care le-am făcut în anii 1950, când au fost create inteligenţa artificială şi cibernetica. Problemele par similare, dar nu sunt. Am avansat, şi acum miza este mai mare. Nu mai este vorba doar despre roboţi cu inteligenţă artificială versus persoane. Este vorba despre inteligenţa artificială care ghidează întregi ecologii.

Cum putem face un ecosistem om-inteligenţă artificială bun, ceva care să nu fie o societate a maşinilor, ci o cultură cibernetică în care putem trăi cu toţii ca oameni – o cultură pe care să o resimţim ca omenească? Nu vrem să gândim la scară mică – de exemplu, să vorbim doar despre roboţi şi despre maşini autonome. Dorim să fie vorba despre o ecologie globală. Ne gândim să fie la scara Skynet. Dar cum am putea face Skynetul să fie de partea oamenilor?

Prima întrebare pe care trebuie să ne-o punem este: care este magia care face inteligenţa artificială prezentă să funcţioneze? Unde greşeşte şi unde nu?

Magia bună este că are ceva numit funcţia de alocare a creditelor. Aceasta îţi permite să iei „neuronii proşti" – funcţii liniare mărunte – şi să-ţi dai seama, într-o reţea mare, care dintre ei fac munca efectivă şi să-i întăreşti. Este o metodă de a face mai inteligente câteva comutatoare aleatorii puse în reţea dându-le *feedback* cu privire la ce funcţionează şi ce nu. Sună simplu, dar matematica pe care se bazează este complicată. Aceasta este magia care face inteligenţa artificială curentă să funcţioneze.

Partea proastă este că, din cauza faptului că acei neuroni mici sunt proşti, lucrurile pe care le învaţă nu se generalizează foarte bine. Dacă o inteligenţă artificială vede ceva ce nu a mai văzut sau dacă lumea se schimbă puţin, inteligenţa artificială este pasibilă de a face o greşeală uriaşă. Nu are nicio idee despre context. Din anumite puncte de vedere, este cât se poate de departe de noţiunea originală a lui Norbert Wiener de cibernetică, pentru că nu este contextualizată; este un mic savant idiot.

Dar să ne imaginăm că îndepărtăm aceste limitări: să ne imaginăm că, în loc să folosim neuroni proşti, folosim neuroni în care au fost integrate cunoştinţe despre lumea reală. Poate că, în loc de neuroni liniari, folosim neuroni care sunt funcţii în fizică şi apoi încercăm să potrivim datele fizice. Sau poate că integrăm multe cunoştinţe despre

oameni şi cum interacţionează ei unul cu celălalt – statisticile şi caracteristicile oamenilor.

Când adăugăm aceste cunoştinţe contextuale şi le conferim şi o funcţie bună de atribuire a creditelor, putem lua datele din observaţii şi folosi funcţia de atribuire a creditelor pentru a întări funcţiile care produc răspunsuri bune. Rezultatul este o inteligenţă artificială care funcţionează foarte bine şi poate generaliza. De exemplu, în rezolvarea problemelor fizice, adesea necesită doar câteva puncte de date zgomotoase pentru a obţine ceva ce este o descriere frumoasă a unui fenomen, pentru că introducem cunoştinţe despre cum funcţionează fizica. Această perspectivă prezintă un contrast uriaş cu inteligenţa artificială normală, care necesită milioane de exemple pentru învăţare şi este foarte sensibilă la zgomot. Adăugând cunoştinţele contextuale corespunzătoare obţinem mult mai multă inteligenţă.

Similar cu cazul sistemelor fizice, dacă vom face neuroni care ştiu multe despre cum învaţă oamenii unul de la altul, atunci vom putea detecta mode omeneşti şi prognoza tendinţele de comportament ale oamenilor surprinzător de precis şi eficient. Această „fizică socială" funcţionează, deoarece comportamentul uman este determinat în aceeaşi măsură de tiparele culturii noastre, cât şi de gândirea individuală şi raţională. Aceste tipare pot fi descrise matematic şi utilizate pentru a face predicţii precise.

Această idee de funcţie de atribuire a creditelor care întăreşte conexiunile dintre neuronii care fac cea mai bună treabă este nucleul inteligenţei artificiale curente. Dacă faci acei neuroni mici mai inteligenţi, inteligenţa artificială devine mai inteligentă. Dar ce s-ar întâmpla dacă am înlocui neuronii cu oameni? Oamenii au multe capacităţi. Ştiu multe despre lume. Pot percepe lucrurile într-un mod în general competent şi omenesc. Ce s-ar întâmpla dacă am avea o reţea de oameni în care am putea întări conexiunile care ajută şi minimaliza conexiunile care nu ajută?

Începe să sune ca o societate sau o companie. Cu toţii trăim într-o reţea socială de oameni. Suntem „întăriţi" când facem lucruri care par să ajute pe toată lumea şi descurajaţi să facem lucruri care nu sunt apreciate. Cultura este rezultatul acestui tip de inteligenţă artificială omenească aplicată problemelor omeneşti; este procesul de a construi structuri sociale întărind conexiunile bune şi penalizându-le pe cele rele. După ce realizăm că putem folosi acest cadru general pentru inteligenţa artificială şi putem crea o inteligenţă artificială umană, întrebarea devine: Care este modul corect de a face asta? Este o idee sigură? Este nebunească?

Eu şi studenţii mei analizăm modul în care oamenii iau decizii din baze de date uriaşe de decizii financiare, decizii de afaceri şi multe

alte tipuri de decizii. Am constatat că oamenii iau adesea decizii într-un mod care imită algoritmii de atribuire a creditelor folosiţi de inteligenţa artificială şi care funcţionează pentru a face comunitatea mai isteaţă. O caracteristică interesantă a acestei activităţi este că analizează o problemă clasică din evoluţie cunoscută sub numele de problema selecţiei de grup. Nucleul acestei probleme este: cum putem selecta cultura în evoluţie, când indivizii sunt cei care se reproduc? Avem nevoie de ceva ce selectează cele mai bune culturi şi cele mai bune grupuri, dar, în acelaşi timp, selectează şi cei mai buni indivizi, pentru că ei reprezintă unităţile care transmit genele.

Când formulăm întrebarea astfel şi parcurgem literatura matematică, descoperim că există o metodă foarte bună. Este numită „distribuţia eşantionului" de tip Thompson, un algoritm matematic utilizat pentru alegerea, dintr-un set de acţiuni posibile cu rezultate necunoscute, acţiunea care maximizează recompensa aşteptată de la acţiuni. Cheia este reprezentată de eşantionarea socială, un mod de a combina dovezile, de a explora şi exploata în acelaşi timp. Are neobişnuita proprietate de a fi simultan cea mai bună strategie atât pentru individ, cât şi pentru grup. Dacă folosim grupul ca bază de selecţie şi grupul este eliminat sau întărit, îi selectăm în acelaşi timp şi pe indivizii de succes. Dacă-i selectăm pe indivizi şi fiecare individ face ce este bine pentru sine, atunci face automat ce este bine pentru grup. Este o aliniere uimitoare a intereselor şi utilităţilor şi oferă observaţii importante în privinţa modului în care cultura se potriveşte în selecţia naturală.

Eşantionarea socială, explicată simplu, înseamnă să te uiţi în jur la acţiunile oamenilor care îţi seamănă, observând ce este popular şi apoi copiind dacă ţi se pare o idee bună. Propagarea ideilor are această funcţie de popularitate la bază, dar adoptarea individuală ţine şi de felul în care ideea funcţionează pentru individ – o atitudine reflexivă. Când combinăm eşantionarea socială şi judecata personală, obţinem un proces superior de luare a deciziilor. Este uimitor, pentru că acum avem o reţetă matematică de a face cu oamenii ce fac toate acele tehnici de inteligenţă artificială cu neuronii proşti ai computerelor. Avem posibilitatea de a alătura oameni pentru a lua decizii mai bune, dată fiind experienţa din ce în ce mai mare.

Aşadar ce se întâmplă în lumea reală? De ce nu facem astfel tot timpul? Ei bine, oamenii sunt pricepuţi la asta, dar există posibilităţi de a o da în bară. Una este prin publicitate, propagandă sau „ştiri false". Există multe moduri de a-i face pe oameni să creadă că ceva este popular atunci când nu este, iar acest lucru distruge utilitatea eşantionării sociale. Modul în care putem crea grupuri de oameni mai inteligenţi, modul în care putem face o inteligenţă artificială umană va funcţiona doar dacă le putem transmite un *feedback* util. Dacă

acţiunile fiecărei persoane au funcţionat pentru ea sau nu, e ceea ce ar trebui să stea la baza acestui proces.

Aceasta este şi cheia mecanismelor de inteligenţă artificială. Ce fac este să analizeze dacă au executat sarcinile corect. Dacă da, plus unu; dacă nu, minus unu. Avem nevoie de acel *feedback* sincer pentru a face acest mecanism omenesc să funcţioneze bine, şi avem nevoie de feluri bune de a şti ce fac alţi oameni pentru a putea evalua corect popularitatea şi probabilitatea de a fi o alegere bună.

Următoarea etapă este construirea acestei funcţii de atribuire a creditelor, această funcţie de *feedback*, pentru oameni, pentru a putea face un ecosistem bun pentru om şi artificial – o organizaţie inteligentă şi o cultură inteligentă. Într-un fel, trebuie să duplicăm câteva observaţii preliminare care au condus la crearea recensământului american, de exemplu – să încercăm să găsim dovezi de bază cu care toată lumea să fie de acord şi să înţelegem astfel că transmiterea cunoştinţei şi culturii poate avea loc într-un mod veritabil şi eşantionarea socială poate funcţiona eficient.

Putem aborda problema creării unei funcţii de atribuire a creditelor în multe contexte. În companii, de exemplu, poate fi făcută cu ecusoane de identificare digitală care descoperă cine este conectat cu cine, pentru a putea evalua tiparul de legături în raport cu rezultatul companiei, zilnic sau săptămânal. Funcţia de atribuire a creditelor se întreabă dacă acele legături au ajutat la rezolvarea problemelor sau la inventarea unor soluţii noi, şi întăreşte legăturile utile. Când poţi obţine acel *feedback* cantitativ – ceea ce este dificil, întrucât cele mai multe lucruri nu se măsoară cantitativ –, atât productivitatea, cât şi rata de inovare din organizaţie pot fi îmbunătăţite semnificativ. Aceasta este, de exemplu, baza metodei de „îmbunătăţire continuă” de la Toyota.

O etapă următoare ar fi să încercăm să facem acelaşi lucru, însă la scară, ceva ce eu numesc construirea unei reţele de încredere pentru date. Poate fi considerat un sistem distribuit, ca internetul, dar cu capacitatea de a măsura cantitativ şi de a comunica astfel calităţile societăţii umane, în acelaşi mod în care recensământul american face o treabă foarte bună informându-ne cu privire la populaţie şi speranţa de viaţă. Deja implementăm exemple prototip de reţele de încredere la scară în mai multe ţări, bazate pe datele şi standardele de măsurare expuse în obiectivele de dezvoltare sustenabilă ale Organizaţiei Naţiunilor Unite.

La orizont avem o viziune pentru cum am putea face umanitatea mai inteligentă construind o inteligenţă artificială umană. Este o viziune compusă din două părţi. Una este reprezentată de datele în care putem avea cu toţii încredere – date care au fost aprobate de o comunitate largă, date pentru care algoritmii sunt cunoscuţi şi

monitorizaţi, asemenea datelor recensământului pe care ne bazăm cu toţii automat ca fiind cel puţin aproximativ corecte. Cealaltă este o evaluare onestă, bazată pe date, despre condiţiile curente. Această a doua parte depinde de disponibilitatea datelor de încredere, prin urmare abia începe să fie dezvoltată. Datele de încredere şi evaluările bazate pe date ale normelor, politicilor şi guvernelor creează împreună o funcţie de atribuire a creditelor care îmbunătăţeşte sănătatea generală şi inteligenţa societăţii.

Exact aici, ştirile false, propaganda şi reclamele se pun în calea înfăptuirii unei inteligenţe mai mari a societăţii. Din fericire, reţelele de încredere ne permit să mergem înainte şi să construim o societate mai rezistentă la problema camerei de rezonanţă, aceste mode, aceste exerciţii de nebunie. Am început să dezvoltăm o modalitate nouă de a stabili măsurile sociale pentru a ne putea ajuta să vindecăm câteva afecţiuni de care suferă societatea de astăzi. Folosim date deschise din toate sursele, încurajând o reprezentare corectă a lucrurilor pe care le aleg oamenii, într-un cadru matematic îngrijit care poate amuţi ecourile şi încercările de manipulare a noastră.

Despre polarizare şi inegalitate

Polarizarea extremă şi segregarea după venit sunt prezente aproape în orice colţ al lumii şi ameninţă să destrame guvernele şi societăţile civile. Din ce în ce mai mult, mass-media devine un traficant de adrenalină împins de la spate de clicurile date pe reclame şi nu mai reuşeşte să raporteze echilibrat evenimentele, într-un discurs raţional, iar degradarea mass-mediei îi năuceşte pe oameni. Nu mai ştiu ce să creadă, putând fi astfel uşor manipulaţi. Există o nevoie reală de a înrădăcina diversele noastre culturi în standarde de încredere şi care au la bază date cu care suntem cu toţii de acord şi de a putea să ştim ce comportamente şi politici funcţionează sau nu.

În convertirea la o societate digitală, am pierdut contactul cu noţiunile tradiţionale de adevăr şi dreptate. Dreptatea era înainte preponderent informală şi normativă. Acum am formalizat-o. În acelaşi timp, aceasta nu mai este la îndemâna celor mai mulţi oameni. Sistemele noastre legale eşuează cum nu eşuau înainte, fix pentru că sunt acum mai formale, mai digitale, mai puţin încorporate în societate.

Ideile despre dreptate sunt foarte diferite în jurul lumii. Unul dintre elementele diferenţiatoare de bază este următorul: tu sau părinţii tăi vă amintiţi când răufăcătorii au venit cu arme şi v-au luat totul? Dacă îţi aduci aminte, atunci atitudinea ta faţă de dreptate este diferită de cea a cititorului obişnuit al acestui eseu. Provii din pătura superioară a societăţii? Sau eşti cineva care a văzut canalele dinăuntru? Perspectiva ta asupra dreptăţii depinde de viaţa pe care ai

trăit-o.

Un test obişnuit pe care-l am pentru cetăţenii americani este următorul: cunoşti pe cineva care are o camionetă? Este cel mai bine vândut autovehicul din Statele Unite şi, dacă nu ştii astfel de oameni, nu eşti în contact cu mai mult de 50% din populaţia Americii. Segregarea fizică stă la baza segregării conceptuale. Cei mai mulţi americani privesc dreptatea, accesul şi corectitudinea foarte diferit de locuitorul obişnuit din Manhattan, să spunem.

Dacă privim tiparele de mobilitate – unde merg oamenii – dintr-un oraş obişnuit, observăm că oamenii din cincimea superioară (familiile care au locuri de muncă la birouri) şi cincimea inferioară (oamenii care trăiesc din şomaj sau ajutor social) nu comunică aproape niciodată. Nu merg în aceleaşi locuri, nu vorbesc despre aceleaşi lucruri. Locuiesc cu toţii în acelaşi oraş, cu numele, dar este ca şi cum ar fi două oraşe diferite – şi aceasta este probabil cea mai importantă cauză a plăgii polarizării din prezent.

Despre bogăţia extremă

Aproximativ două sute dintre cei mai bogaţi oameni din lume au promis că vor da mai mult de jumătate din averea lor fie în timpul vieţii, fie prin testament, creând astfel o pluralitate de voci în spaţiul fundaţiilor[59]. Bill Gates este probabil cel mai cunoscut exemplu. A decis că, dacă guvernul nu o face, o va face el. Vrei plase de ţânţari? Le face el. Vrei medicamente antivirale? Le face el. Facem ca diverse persoane interesate să acţioneze ca fundaţii dedicate binelui public, însă acestea au versiuni diferite pentru ceea ce consideră a fi binele public. Această diversitate de obiective a creat o mare parte din frumuseţea lumii contemporane. Acţiuni întreprinse în afara guvernelor de organizaţii precum Fundaţia Ford şi Fundaţia Sloan, care pariază pe lucruri pe care nimeni altcineva nu ar paria, au schimbat lumea în bine.

Bineînţeles, aceşti miliardari sunt oameni, cu metehne omeneşti, şi nu totul este neapărat aşa cum ar trebui să fie. Pe de altă parte, aceeaşi situaţie a existat şi când au început să se construiască căi ferate. Unii oameni au făcut averi uriaşe. Mulţi au dat faliment. Noi, oamenii obişnuiţi, ne-am ales cu căile ferate. Asta e bine. La fel şi cu energia electrică; la fel cu multe tehnologii noi. Există un proces de agitaţie care aruncă pe cineva sus şi, mai târziu, îi duce pe el sau pe moştenitorii săi jos. Bulele de bogăţie extremă erau o caracteristică a sfârşitului secolului al XIX-lea şi Începutului secolului XX, când au fost inventate motoarele cu abur, căile ferate şi luminile electrice. Averile pe care le-au creat dispăreau complet în două-trei generaţii.

Dacă Statele Unite ar fi fost ca Europa, mi-aş fi făcut probleme. Ce constatăm în Europa este că aceleaşi familii au ţinut de averi sute de

ani, deci sunt înrădăcinate nu doar ca avuţie, dar şi în sistemul politic şi în alte domenii. Dar, până acum, SUA a evitat acest fel de sistem ereditar de clasă. Bogăţia extremă nu a rămas de neatins, ceea ce este bine. Nu trebuie să fie de neatins. În cazul în care câştigi la loterie, primeşti miliardul de dolari, dar nepoţii tăi trebuie să muncească pentru a trăi.

Despre inteligenţa artificială şi societate

Oamenii se tem de inteligenţa artificială. Poate că aşa şi trebuie. Dar e bine să realizeze că inteligenţa artificială se hrăneşte cu date. Fără date, inteligenţa artificială nu reprezintă nimic. Nu trebuie să fii cu ochii pe inteligenţa artificială; în schimb trebuie să fii cu ochii pe ce consumă şi ce face. Cadrul reţelei de încredere pe care l-am implementat, cu ajutorul unor naţiuni din Uniunea Europeană şi din alte zone, este unul în care putem avea algoritmii noştri, putem avea inteligenţa noastră artificială, dar putem să vedem ce a intrat şi ce a ieşit din proces pentru a ne putea întreba: este aceasta o decizie discriminatorie? Asta vrem ca oameni? Sau este ceva puţin cam bizar?

Cea mai grăitoare analogie este că agenţiile de reglementare, birocraţiile şi părţile din guvern se aseamănă foarte mult cu inteligenţa artificială: preiau regulile pe care le numesc legi sau reglementări şi adaugă la ele date guvernamentale, apoi iau decizii care ne afectează vieţile. Partea care este rea din sistemul curent este că avem foarte puţin control asupra acestor departamente, agenţii de reglementare şi birocraţii. Singurul control pe care-l avem este votul – oportunitatea de a alege o altă persoană. Trebuie să facem controlul asupra birocraţiilor mult mai fin. Trebuie să înregistrăm datele care au intrat în fiecare decizie şi să analizăm rezultate în funcţie de diversele părţi interesate – aşa cum se intenţiona iniţial să facă legislaturile alese.

Dacă avem datele care intră şi ies din fiecare decizie, putem să ne întrebăm cu uşurinţă: este acesta un algoritm bun? Face această inteligenţă artificială lucruri pe care noi, oamenii, le considerăm etice? Această abordare cu oameni în ciclul de funcţionare este numită „algoritm deschis”; adică putem vedea ce folosesc inteligenţele artificiale ca intrări şi ce decid folosind acele date de intrare. Dacă vedem aceste două lucruri, vom şti dacă facem bine sau nu. Se pare că nu este greu de făcut. Dacă vom controla datele, vom controla inteligenţa artificială.

Un lucru pe care adesea oamenii omit să-l menţioneze este că toate grijile legate de inteligenţa artificială sunt identice cu grijile legate de guvernele contemporane. În cea mai mare parte a proceselor guvernamentale – sistemul de justiţie etc. – nu există date fiabile despre ce fac şi în ce situaţii anume. Cum putem şti dacă instanţele de

judecată sunt imparţiale dacă nu ştim intrările şi ieşirile? Aceeaşi problemă ia naştere în legătură cu sistemele de inteligenţă artificială şi poate fi abordată în acelaşi fel. Avem nevoie de date de încredere pentru a trage la răspundere guvernele curente cu privire la datele folosite şi deciziile luate în baza acestora, iar inteligenţa artificială nu trebuie să se sustragă de la această regulă.

Inteligenţa generală din generaţia viitoare

Algoritmii prezenţi de învăţare automată folosiţi de inteligenţele artificiale sunt, la bază, simpli ca bună ziua. Funcţionează, dar funcţionează prin forţă brută, deci au nevoie de sute de milioane de eşantioane. Funcţionează pentru că poţi aproxima orice cu multe bucăţi simple. Aceasta este o observaţie-cheie a cercetării prezente din domeniul inteligenţei artificiale – şi anume, dacă foloseşti învăţarea prin întărire pentru *feedbac*kul de atribuire a creditelor, poţi face acele bucăţele să aproximeze orice funcţie arbitrară doreşti.

Dar folosirea unor funcţii greşite pentru a lua decizii înseamnă că abilitatea inteligenţei artificiale de a lua decizii bune nu va fi generalizată. Dacă-i dăm inteligenţei artificiale intrări noi, diferite, poate lua decizii complet nerezonabile. Sau, dacă situaţia se schimbă, atunci trebuie să o reinstruim. Există tehnici amuzante pentru a găsi „spaţiile nule” în aceste sisteme de inteligenţă artificială. Acestea sunt intrări despre care inteligenţa artificială crede că sunt exemple valabile pentru ceea ce a fost antrenată să recunoască (de exemplu feţe, pisici etc.), dar pentru un om sunt exemple nebuneşti.

Inteligenţa artificială curentă face statistici descriptive într-un mod care nu este ştiinţific şi ar fi aproape imposibil de transformat în ştiinţă. Pentru a construi sisteme robuste, trebuie să cunoaştem ştiinţa din spatele datelor. Sistemele pe care le văd ca inteligenţe artificiale de nouă generaţie iau naştere din abordarea bazată pe ştiinţă: dacă urmează să creezi o inteligenţă artificială care să rezolve o problemă fizică, atunci trebuie să integrezi în ea legile fizicii drept funcţii descriptive, în locul acelor neuroni proşti. De exemplu, ştim că fizica foloseşte funcţii polinomiale, sinusoidale şi exponenţiale, deci acestea ar trebui să fie funcţiile de bază, şi nu mici neuroni liniari. Folosind aceste funcţii de bază mai potrivite, avem nevoie de mult mai puţine date, putem face faţă unui zgomot mult mai mare şi obţinem rezultate mult mai bune.

Ca în exemplul cu fizica, dacă dorim să construim o inteligenţă artificială care să lucreze cu comportamentul uman, trebuie să construim proprietăţile statistice ale reţelelor omeneşti în algoritmi de învăţare automată. Când înlocuim neuronii proşti cu neuroni care conţin câteva elemente de bază despre comportamentul uman, putem identifica tendinţe cu foarte puţine date şi putem face faţă unui nivel

uriaş de zgomot.

Faptul că oamenii au o înţelegere „intuitivă” pe care o folosesc în majoritatea problemelor sugerează ceea ce eu numesc strategia umană: societatea umană este o reţea la fel ca reţelele neuronale instruite pentru învăţare profundă, dar „neuronii” din societăţile omeneşti sunt mult mai inteligenţi. Tu şi eu avem puteri descriptive surprinzător de generale pe care le folosim pentru a înţelege o gamă variată de situaţii şi putem recunoaşte ce conexiuni trebuie întărite. Asta înseamnă că ne putem configura reţelele sociale să funcţioneze mult mai bine şi, posibil, să învingă toată acea inteligenţă artificială bazată pe maşini la propriul ei joc.

Capitolul 20

Transformarea invizibilului în vizibil:

Hans Ulrich Obrist

Hans Ulrich Obrist este director artistic la Galeria Serpentine din Londra şi autorul cărţilor Ways of
Curating *şi* Lives of the Artists, Lives of the Architects.

„URGENT! URGENT!", ţipa copia cc a unui e-mail, unul dintre zecile de e-mailuri pe care le-am primit când am pornit telefonul în zona de bagaje din aeroportul Malpensa după zborul lung de la JFK. „Vizionarul american John Brockman ajunge în această dimineaţă la Grand Hotel Milan. TREBUIE, repet, TREBUIE să-i faci o vizită". E-mailul era semnat HUO.

Cu o seară înainte, stând în sala de aşteptare de la JFK, avusesem ideea strălucită să-i scriu prietenului şi colaboratorului meu de mult timp, care locuieşte în Londra, peripateticul curator de artă Hans Ulrich Obrist (pe care toată lumea îl ştie că HUO), şi să-l întreb dacă există vreo persoană în Milano pe care ar trebui să o cunosc.

Când m-am instalat la hotel, telefonul a început să zbârnâie şi o procesiune de artişti, designeri şi arhitecţi italieni au sunat să-mi solicite o întrevedere, printre care şi

Enzo Mari, artistul modernist şi designerul de mobilă; Alberto Garutti, ale cărui strategii estetice au inspirat un dialog între arta contemporană, spectator şi spaţiul public; şi designerul de modă Miuccia Prada, care „vă solicită prezenţa la ceai în această după-amiază la sediul Prada". Şi astfel, mulţumită lui HUO, „marele gânditor american vizionar", obosit după zbor, a trecut cu paşi grei de prima zi a sa în Milano, în noiembrie 2011.

HUO este sui generis: trăieşte o zi de douăzeci şi patru de ore, dormind (cred) pe apucate şi angajând asistenţi cu normă întreagă, care lucrează ture de opt ore şi îi sunt disponibili 24 de ore pe zi, 7 zile pe săptămână. Într-o perioadă recentă de doi ani a vizitat galerii de artă în China sau în India patruzeci de weekenduri în fiecare an – plecând din Londra joia şi revenind la birou lunea. Anul trecut, din nou, ArtReview l-a clasificat din nou ca fiind numărul unu pe lista lor anuală „Power 100".

Recent, am colaborat într-o comisie în timpul evenimentului Serpentine „GUEST, GHOST, HOST: MACHINE!", care a avut loc în noul sediu al Primăriei Londrei. Ni s-au alăturat Venki Ramakrishnan, Jaan Tallinn şi Andrew Blake, director de cercetare la Institutul Alan Turing. Evenimentul se subscria misiunii lui HUO de a aduce laolaltă arta şi ştiinţa: „Curatorul nu mai este înţeles pur şi simplu ca o persoană care umple un spaţiu cu obiecte",

spune el, „ci şi ca persoana care pune în contact sfere culturale diferite, inventează noile caracteristici de afişare şi face legăturile necesare care permit întâlniri şi rezultate neaşteptate”.

În introducerea celei de-a doua ediţii a cărţii sale *Understanding Mediay* Marshall McLuhan a observat capacitatea artei de a „anticipa progresele sociale şi tehnologice ale viitorului”. Arta este „un sistem de alarmă timpuriu” care ne arată noile dezvoltări ce vor urma şi ne permite să „ne pregătim să le facem faţă… Arta ca mediu de radar preia funcţia de instruire perceptuală indispensabilă”.

În 1964, când a fost publicată pentru prima dată cartea lui McLuhan, artistul N-am June Paik tocmai îşi construia *Robot K-456* pentru a experimenta cu tehnologiile care vor urma să influenţeze societatea. Lucrase şi înainte cu televiziunea, provocând consumul pasiv al telespectatorului, şi, mai târziu, a creat artă cu difuzări live prin satelit, utilizând noile medii mai puţin pentru divertisment şi mai mult pentru a evidenţia capacităţile sale poetice şi interculturale (care sunt încă în mare parte neutilizate în prezent). Paik-ii vremurilor noastre, bineînţeles, lucrează acum cu internetul, imagini digitale şi inteligenţă artificială. Lucrările şi gândurile lor sunt din nou un sistem de alarmă timpuriu pentru dezvoltările care vor urma.

În calitate de custode, activitatea mea zilnică constă în a reuni opere de artă diverse pentru a conecta culturi diferite. De la începutul anilor 1990, organizez conversaţii şi întâlniri cu practicanţi ai unor diverse discipline, pentru a putea trece dincolo de reticenţa generală de acumulare a cunoştinţelor. Deoarece eram interesat să aud ce au de spus artiştii despre inteligenţa artificială, am organizat recent mai multe conversaţii între artişti şi ingineri.

Motivul pentru care privesc cu interes inteligenţa artificială este că două dintre cele mai importante întrebări ale zilelor noastre sunt „Când de capabilă va deveni inteligenţa artificială?” şi „Ce pericole se pot naşte din ea?” Aplicările sale incipiente ne influenţează deja viaţa de zi cu zi în moduri care sunt mai mult sau mai puţin recognoscibile. Există un impact în creştere asupra multor aspecte ale societăţii noastre, dar, dacă acest lucru este, în general, benefic sau malign este încă nesigur.

Mulţi artişti contemporani urmăresc îndeaproape aceste dezvoltări. Articulează diverse îndoieli despre promisiunile inteligenţei artificiale şi ne amintesc să nu asociem termenul de „inteligenţă artificială” numai cu rezultate pozitive. La discuţiile curente despre inteligenţa artificială, artiştii contribuie cu perspectivele lor speciale şi cu concentrarea pe întrebări legate de crearea imaginilor, creativitate şi utilizarea programării ca instrument artistic.

Conexiunile profunde dintre ştiinţă şi artă au fost observate deja

de regretatul Heinz von Foerster, unul dintre arhitecţii ciberneticii, care a lucrat cu Norber Wiener de la jumătatea anilor 1940, iar în anii 1960 a fondat domeniul ciberneticii secundare, în care observatorul este înţeles ca făcând parte din sistem, şi nu ca o entitate externă. L-am cunoscut bine pe von Foerster şi într-una dintre multele conversaţii pe care le-am purtat mi-a comunicat perspectiva sa asupra relaţiei dintre artă şi ştiinţă:

> *Întotdeauna am perceput arta şi ştiinţa ca pe nişte domenii complementare. Nu trebuie să uităm că un om de ştiinţă este în anumite privinţe şi un artist. Inventează o tehnică nouă şi o descrie. Foloseşte limbajul ca un poet sau ca autorul unui roman poliţist şi îşi descrie rezultatele. Din punctul meu de vedere, un om de ştiinţă trebuie să lucreze artistic dacă vrea să-şi comunice cercetările. Este evident că vrea să le comunice şi să vorbească cu alţii. Un om de ştiinţă inventează obiecte noi şi problema este cum vor fi descrise. În toate aceste privinţe, ştiinţa nu este foarte diferită de artă.*

Când l-am întrebat cum defineşte cibernetica, von Foerster a răspuns:

> *Substanţa învăţăturilor noastre trase din cibernetică este să gândim circular: A conduce la B, B la C, dar C se întoarce la A. Astfel de argumente nu sunt liniare, ci sunt circulare. Contribuţia semnificativă a ciberneticii la gândirea noastră este acceptarea argumentelor circulare. Asta înseamnă că trebuie să privim procesele circulare şi să înţelegem care sunt circumstanţele în care ia naştere un echilibru, deci o structură stabilă.*

Astăzi, acolo unde algoritmii de inteligenţă artificială sunt aplicaţi la sarcinile zilnice, ne putem întreba cum este inclus factorul uman în aceste tipuri de procese şi ce roluri pot juca arta şi creativitatea în raport cu ei. Astfel există niveluri diferite de gândire despre explorarea relaţiei dintre inteligenţă artificială şi artă.

Prin urmare, ce au de spus artiştii contemporani despre inteligenţa artificială?

Prostia artificială

Hito Steyerl, o artistă care lucrează cu filme documentare şi experimentale, ia în considerare două aspecte-cheie pe care trebuie să le avem în minte când ne gândim la consecinţele inteligenţei artificiale pentru societate. În primul rând, aşteptările pentru aşa-numita inteligenţă artificială, spune ea, sunt adesea exagerate, iar substantivul „inteligenţă” ne induce adesea în eroare; pentru a contracara acest efect, foloseşte termenul de „prostie artificială”. În al

doilea rând, scoate în evidenţă faptul că programatorii fac acum algoritmii software invizibili să fie vizibili prin imagini, dar, pentru a înţelege şi a interpreta mai bine aceste imagini, trebuie să aplicăm expertiza artiştilor.

Steyerl a lucrat cu computerele mulţi ani şi operele ei recente de artă au explorat tehnicile de supraveghere, roboţii şi jocuri pe calculator, precum în *How Not to be Seen* (2013), despre tehnologii de imagistică digitală, sau *Hell YeahWeFuck Die* (2017), despre instruirea roboţilor în obiectivul echilibrului, care încă este dificilă pentru ei. Dar, pentru a explica noţiunea ei de prostie artificială, Steyerl face trimitere la un fenomen mai general, cum este utilizarea acum generalizată a „boţilor" Twitter, observând în conversaţia noastră:

> *A fost şi încă este o unealtă foarte populară în alegeri să angajezi armate Twitter să încline opinia publică şi să devieze hashtag-urilz populare etc. Aceasta este o inteligenţă artificială de un nivel foarte, foarte scăzut. Înseamnă două sau poate trei linii de script. Nu este deloc sofisticat. Însă implicaţiile sociale ale acestui tip de prostie artificială, cum o numesc eu, sunt deja monumentale în politica globală.*

După cum a fost observat pe scară largă, acest tip de tehnologie a fost văzut în multe postări Twitter automate înainte de alegerile prezidenţiale din 2016, în Statele Unite, şi, de asemenea, cu puţin timp înainte de votul Brexit. Dacă până şi o tehnologie de inteligenţă artificială atât de simplă cum sunt aceşti „boţi" influenţează deja politica, ia naştere altă întrebare urgentă: „Cât de puternice vor fi tehnicile mult mai avansate din viitor?"

Vizibil/invizibil

Artistul Paul Klee vorbea adesea despre artă ca „a face invizibilul vizibil". În tehnologia informaţiei, majoritatea algoritmilor lucrează invizibil, în fundal; rămân inaccesibili în sistemele pe care le folosim zi de zi. Dar, în ultima perioadă, s-a observat o revenire interesantă a vizualului în învăţarea automată. Modurile în care algoritmii de învăţare automată ai inteligenţei artificiale prelucrează datele au fost făcuţi vizibili prin aplicaţii precum Deep Dream de la Google, în care procesul de recunoaştere computerizată a tiparelor este vizualizat în timp real. Aplicaţia arată cum algoritmul încearcă să potrivească formele de animale cu orice intrare dată. Există multe alte programe de vizualizare bazate pe inteligenţă artificială care, în felul lor, fac şi ele „invizibilul vizibil". Dificultatea din percepţia publicului larg a acestor imagini este, după părerea lui Steyerl, că aceste tipare vizuale sunt privite fără critică drept reprezentări realiste şi obiective ale procesului automat. Spune despre estetica unor astfel de vizualizări:

Pentru mine, acest lucru dovedeşte că ştiinţa a devenit un subgen al istoriei artei… Avem acum multe tipare abstracte de computer care pot arăta ca o pictură de Paul Klee sau un Mark Rothko sau tot felul de alte abstracţiuni pe care le ştim din istorie. Singura diferenţă, cred eu, este că, în gândirea ştiinţifică curentă, sunt percepute ca reprezentări ale realităţii, aproape ca imaginile unui documentar, în timp ce în istoria artei există o înţelegere foarte nuanţată a diverselor tipuri de abstractizări.

Ceea ce caută ea este o înţelegere mai profundă a imaginilor generate de computer şi diversele forme estetice pe care le utilizează acestea. Evident, nu sunt generate cu scopul explicit de a respecta o anumită tradiţiei estetică. Inginerul IT Mike Tyka, într-o conversaţie purtată cu Steyerl, explica funcţiile acestor imagini:

Sistemele de învăţare profundă, în special cele vizuale, sunt de fapt inspirate de nevoia de a şti ce se întâmplă în cutia neagră. Scopul lor este să proiecteze acele procese înapoi în lumea reală.

Cu toate acestea, aceste imagini au implicaţii estetice şi valori care trebuie luate în considerare. Se poate spune că, în timp ce programatorii folosesc imagini pentru a ne ajuta să înţelegem mai bine algoritmii programelor, avem nevoie de cunoştinţele artiştilor pentru a înţelege mai bine formele estetice ale inteligenţei artificiale. Cum a evidenţiat Steyerl, astfel de vizualizări sunt în general înţelese ca reprezentări „adevărate" ale proceselor, dar trebuie să fim atenţi la esteticile lor corespunzătoare şi la implicaţiile acestora, care trebuie văzute critic şi analitic.

În 2017, artistul Trevor Paglen a creat un proiect de a face vizibili aceşti algoritmi invizibili ai inteligenţei artificiale. In Sight *Machine*, a filmat o reprezentaţie live a cvartetului Kronos şi a procesat imaginile rezultate cu diverse programe software folosite pentru detectare facială, identificarea obiectelor şi chiar ghidarea rachetelor balistice. A proiectat rezultatul acestor algoritmi, în timp real, înapoi pe ecrane în spatele scenei. Demonstrând cum interpretează performanţa muzicienilor diversele programe, Paglen a arătat că algoritmii inteligenţei artificiale sunt determinaţi întotdeauna de seturi de valori şi interese pe care apoi le manifestă şi reiterează, trebuind astfel interogaţi critic. Contrastul semnificativ dintre algoritmi şi muzică ridică şi problema relaţiilor dintre percepţia tehnică şi cea omenească.

Computerele, ca instrument de creativitate, nu-l pot înlocui pe artist

Rachel Rose, o artistă video care se gândeşte la problemele puse de inteligenţa artificială, foloseşte tehnologia computerelor în crearea

operelor sale. Filmele sale conferă privitorului o experienţă a materialităţii prin intermediul imaginii în mişcare. Foloseşte colajul şi stratificarea materialului pentru a manipula sunetul şi imaginea, iar procesul de editare este probabil cel mai important aspect al activităţii sale.

De asemenea, vorbeşte despre importanţa procesului de luare a deciziilor în activitatea sa. Pentru ea, procesul artistic nu respectă un tipar raţional. Într-o conversaţie pe care am avut-o, împreună cu inginerul Kenric McDowell, de la Google Cultural Institute, a explicat această idee citând o povestioară din cartea regizorului de teatru Peter Brook *The Empty Space / Spaţiul gol*, din 1968. Când Brook a conceput decorul pentru spectacolul său *Furtuna*, la sfârşitul anilor 1960, a început prin a face o grădină japoneză, dar apoi conceptul a evoluat, devenind o cutie albă, o cutie neagră, o scenografie realistă şi aşa mai departe. Şi, în cele din urmă, s-a întors la ideea sa iniţială. Brook scrie că a fost şocat de faptul că a pierdut o lună întreagă doar pentru a sfârşi unde începuse. Dar acest lucru arată că procesul artistic creativ este o succesiune a cărei fiecare etapă se construieşte pe cea anterioară şi, în cele din urmă, ajunge la o concluzie impredictibilă, legată de propriul proces artistic de luare a deciziilor:

> *Pentru mine, este foarte diferit de învăţarea automată, pentru că la baza fiecărei decizii se află acest sentiment adânc care vine din fiinţa umană, care are de-a face cu empatia, care are de-a face cu comunicarea, care are de-a face cu întrebări cu privire la propria noastră mortalitate pe care doar un om le poate pune.*

Această idee evidenţiază diferenţa fundamentală dintre orice producţie artistică umană şi aşa-zisa creativitate a computerelor. Rose vede inteligenţa artificială mai mult ca pe o cale de a crea instrumente mai bune pentru oameni:

> *Îmi imaginez că învăţarea automată îl poate ajuta pe artist nu în dezvoltarea unei subiectivităţi independente – cum ar fi compunerea unui poem sau crearea unei imagini –, ci în umplerea golurilor care au de-a face cu munca manuală, cum face Photoshopul cu diversele instrumente pe care le poţi utiliza.*

Şi, deşi aceste instrumente poate nu par spectaculoase, spune ea, „ar putea avea o influenţă importantă asupra artei”, pentru că pun la dispoziţia artiştilor mai multe posibilităţi în activitatea lor creativă.

McDowell a adăugat că şi el crede că există aşteptări false de la inteligenţa artificială. „Am observat”, spune el, „că ideea unui computer care să facă toate lucrurile pe care le facem şi noi este oarecum magică”. A continuat: „există aproape un fel de oglindă demonică în care privim şi vrem ca imaginea să scrie un roman, să

facă un film – vrem să predăm cumva această putere pe care o avem". În schimb, lucrează la proiecte în care oamenii colaborează cu maşinile. Unul dintre scopurile prezente ale cercetării din domeniul inteligenţei artificiale este să găsească noi mijloace de interacţiune între oameni şi software. Şi arta, se poate spune, trebuie să joace un rol-cheie în această întreprindere, din moment ce se concentrează pe subiectivitatea noastră şi pe aspecte omeneşti esenţiale, cum ar fi empatia şi mortalitatea.

Cibernetică/artă

Suzanne Treister este o artistă a cărei activitate dintre anii '2009 şi 2011 serveşte drept exemplu pentru ceea ce se întâmplă la intersecţia dintre tehnologiile noastre curente, arte şi cibernetică. Treister a fost o pioniera în arta digitală încă din anii 1990, inventând, de exemplu, jocuri video imaginare şi pictând capturi de ecran din ele. În proiectul său *HEXEN2. O*, a analizat faimoasele conferinţe de cibernetică Macy, care au fost organizate între 1946 şi 1953, în New York, de ingineri şi cercetători din domeniul ştiinţelor sociale, pentru a uni ştiinţele şi pentru a dezvolta o teorie universală a mecanismelor minţii.

În proiectul său, a creat treizeci de lucrări foto-text despre participanţii la conferinţa sa (printre care se numărau şi Wiener şi von Foerster), a inventat cărţi de tarot şi a făcut un video bazat pe un montaj fotografic al unei „şedinţe cibernetice de spiritism". Participanţii la conferinţă se pot vedea stând în jurul unei mese rotunde, ca în şedinţele de spiritism, în timp ce câteva dintre afirmaţiile lor despre cibernetică se aud pe colajul audio – cunoştinţe raţionale şi superstiţii combinate. De asemenea, a observat că anumiţi participanţi la şedinţă lucrau pentru armată; astfel, aplicarea ciberneticii putea fi privită ambivalent, chiar şi atunci, ca o luptă între cunoştinţele pure şi utilizarea ei în controlul exercitat de stat.

Dacă privim lucrarea lui Treister despre participanţii la conferinţa Macy, vedem că nu a fost inclus niciun artist vizual. Un dialog între artişti şi oamenii de ştiinţă va fi fructuos în discuţiile viitoare şi este puţin uimitor că nu şi-au dat seama de acest lucru la momentul respectiv, dat fiind interesul profund al lui Foerster pentru artă. Şi-a amintit într-una dintre conversaţiile noastre cum relaţia lui cu domeniul îşi avea rădăcinile chiar în copilărie:

> *Am crescut într-o familie de artişti. Adesea eram vizitaţi de poeţi, filosofi, pictori şi sculptori. Arta făcea parte din viaţa mea. Mai târziu, a început să mă intereseze fizica, întrucât se pare că aveam talent. Dar am rămas întotdeauna conştient de importanţa artei pentru ştiinţă.*

Nu era o diferenţă mare pentru mine. Pentru mine, ambele aspecte ale vieţii erau foarte asemănătoare şi accesibile. Ar trebui să le vedem ca pe unul singur. Şi un artist trebuie să reflecteze asupra lucrărilor sale. Trebuie să se gândească la gramatică şi la limbaj. Un pictor trebuie să ştie cum să-şi gestioneze culorile. Gândiţi-vă cât de intens erau cercetate culorile în timpul Renaşterii. Voiau să ştie cum un anumit pigment putea fi amestecat cu alţii pentru a obţine o anumită nuanţă de roşu sau albastru. Chimiştii şi pictorii colaborau îndeaproape. Cred că împărţirea artificială între ştiinţe şi arte este greşită.

Deşi pentru von Foerster relaţia dintre artă şi ştiinţă a fost întotdeauna clară, pentru vremurile noastre această legătură nu este făcută încă. Există multe motive pentru a multiplica legăturile. Gândirea critică a artiştilor ar fi benefică în ceea ce priveşte pericolele inteligenţei artificiale, deoarece ei ne atrag atenţia asupra întrebărilor pe care le consideră esenţiale din perspectiva proprie. Cu progresele făcute de învăţarea automată, artiştii au la dispoziţie instrumente noi pentru activitatea lor. Şi, pe măsură ce algoritmii inteligenţei artificiale sunt făcuţi vizibili prin imagini artificiale în moduri noi, expertiza şi cunoştinţele vizuale critice ale artiştilor vor fi folosite. Multe dintre aceste întrebări-cheie y despre inteligenţa artificială sunt de natură filosofică şi pot primi răspuns doar dintr-o perspectivă holistică. Modul în care prind viaţă în rândurile artiştilor aventuroşi va merita urmărit.

Simularea unor lumi

În cea mai mare parte, lucrările artiştilor contemporani au fost meditaţii întruchipate ale impactului inteligenţei artificiale asupra unor întrebări existenţiale despre sine şi interacţiunile noastre viitoare cu entităţile nonumane. Puţini însă au privit tehnologiile şi inovaţiile inteligenţei artificiale ca pe materia primă a creaţiilor lor şi le-au sculptat în propria viziune. O excepţie este reprezentată de artistul Ian Cheng, care a mers până în punctul în care a construit lumi întregi de fiinţe artificiale cu diverse grade de conştiinţă şi inteligenţă. A numit aceste lumi „simulări live". Trilogia sa *Emisarii* (2015 – 2017) este plasată într-o lume ficţională postapocaliptică cu floră şi faună, în care animale şi creaturi controlate de inteligenţa artificială explorează peisajul şi interacţionează unele cu altele. Cheng foloseşte o grafică avansată, dar le-a programat cu multe erori şi imperfecţiuni, creând astfel o atmosferă futuristă şi anacronică în acelaşi timp. Prin trilogia sa, care cartografiază un istoric al conştiinţei, pune întrebarea „ce este o simulare?"

În timp ce majoritatea lucrărilor artistice ce utilizează ultimele

tehnologii din domeniul inteligenţei artificiale se inspiră în special din învăţarea automată, simulările live ale lui Chen o iau pe altă cale. Protagoniştii şi poveştile care sunt îmbinate cu fiecare simulare episodică a *Emisarilor* folosesc sistemele logice şi regulile complexe ale inteligenţei artificiale. Ce este profund în scenele sale care evoluează în mod continuu este că complexitatea ia naştere nu din dorinţele/ acţiunile unui singur actor sau zeu artificial, ci din constelaţiile, coliziunile şi evoluţia constantă în simbioza dintre entităţi. Acest lucru dă naştere unor rezultate neaşteptate şi situaţii nesfârşite, care nu pot fi cunoscute – nu poţi experimenta niciodată acelaşi moment dacă vezi repetat lucrările sale.

Cheng a susţinut o conferinţă la maratonul Serpentine „GUEST, GHOST, HOST: MACHINE!", cu programatorul Richard Evans, care a proiectat recent Versu, o platformă bazată pe inteligenţă artificială pentru jocuri interactive de narare a unei poveşti. Lucrările lui Evans scot în evidenţă interacţiunea socială a personajelor din joc, care reacţionează într-un spectru de comportamente posibile la alegerile făcute de jucătorii umani. În conversaţia lor, Evans a spus că un punct de plecare pentru proiectul său a fost faptul că primele jocuri de simulare, cum ar fi The Sims, nu au luat în considerare suficient importanţa practicilor sociale. Protagoniştii simulaţi din jocuri se comportă adesea în moduri care nu corespund comportamentelor umane reale. Cunoaşterea practicilor sociale limitează posibilităţile de acţiune, dar este necesară pentru a înţelege sensul acţiunilor noastre – ceea ce-l interesează pe Cheng pentru propriile sale simulări. Cu cât sunt mai bine determinaţi parametrii acţiunilor în anumite circumstanţe într-o simulare pe computer, cu atât este mai interesant pentru Cheng să experimenteze cu schimbări individuale şi specifice. I-a spus lui Evans: „Consider că, în măsura în care am avea o inteligenţă artificială cu o capacitate mai bună de a răspunde contextelor sociale, dacă ajustăm o calitate obţinem ceva artistic şi frumos".

De asemenea, Cheng priveşte activitatea programatorilor şi simulărilor cu inteligenţă artificială la bază drept crearea unor instrumente noi şi sofisticate pentru experimentarea cu parametrii activităţilor noastre sociale de zi cu zi. În acest mod, implicarea artiştilor în domeniul inteligenţei artificiale va conduce la tipuri noi de experimente deschise în artă. Astfel de posibilităţi sunt – asemenea capacităţilor crescute ale inteligenţei artificiale în general – încă în viitor. Recunoscând că aceasta este o tehnologie experimentală, incipientă, foarte departe de viziunile apocaliptice despre inteligenţe artificiale superinteligente care vor cuceri lumea, Cheng îşi umple simulările cu avataruri prozaice, cum ar fi nişte globule microbiene ciudate, câini şi morţi vii.

Astfel de discuţii între artişti şi ingineri, bineînţeles, nu sunt complet noi. În anii 1960, inginerul Billy Klüver a reunit artişti cu ingineri într-o serie de evenimente, iar în 1967 a pus bazele programului „Experimente în artă şi tehnologie", cu Robert Rauschenberg şi alţii. În Londra, aproximativ în acelaşi timp, Barbara Steveni şi John Latham, din Artist Placement Group, au dus lucrurile cu un pas mai departe, afirmând că ar trebui să existe artişti angajaţi la fiecare companie şi în fiecare guvern. Astăzi aceste modele istorice pline de inspiraţie pot fi aplicate în domeniul inteligenţei artificiale. Pe măsură ce inteligenţa artificială ajunge să fie implicată din ce în ce mai mult în vieţile noastre de zi cu zi, crearea unui spaţiu care este nondeterminist şi nonutilitar în pluralitatea sa de perspective şi diversitatea de înţelegeri va fi cu siguranţă esenţială.

Capitolul 21
Inteligenţele artificiale versus copiii de patru ani

Alison Gopnik

Alison Gopnik este specialistă în psihologia dezvoltării la UC Berkeley. Printre cărţile sale se numără The Philosophical Baby *şi, mai recent,* The Gardener and the Carpenter: What the New Science of Child Development Tells Us About the Relationship Between Parents and Children.

Alison Gopnik este lider internaţional în domeniul învăţării şi dezvoltării copiilor şi a fost una dintre fondatoarele domeniului „teoriei minţii". A vorbit despre creierul de copil ca fiind un „puternic computer de învăţare", poate din experienţă personală. Copilăria ei din Philadelphia a fost un exerciţiu în dezvoltarea intelectuală. „Alte familii îşi duceau copiii să vadă filmul Sunetul muzicii sau Carusel; noi am văzut Fedra lui Racine şi Sfârşitul jocului de Samuel Beckett", îşi amintea ea. „Familia noastră citea cu voce tare romanul Joseph Andrewsscris de Henry Fielding în secolul al XVIII-lea, în jurul focului de tabără, când mergeam la camping".

Recent, a invocat modelele bayesiene ale învăţării automate pentru a explica abilitatea excepţională a preşcolarilor de a trage concluzii despre lumea din jurul lor fără a beneficia de seturi de date eronate. „Cred că bebeluşii şi copiii sunt mai conştienţi decât suntem noi, adulţii", spunea ea. „Se pricep de minune să absoarbă multă informaţie din multe surse diferite deodată". I-a numit pe bebeluşi şi pe copiii tineri „departamentul de cercetare şi dezvoltare al speciei umane". Asta nu înseamnă că este rece cu

ei, ca şi cum ar fi nişte simple animale de laborator. Copiii par să se simtă foarte bine în compania ei şi printre jucăriile luminoase şi zgomotoase din laboratorul său de la Berkeley. Ani întregi după ce propriii săi copii au trecut de vârsta jucăriilor, ea le-a păstrat în birou.

Investigaţiile sale în procesul nostru de învăţare şi paralelele făcute cu metodele de învăţare profundă ale inteligenţei artificiale continuă. „Se pare că este mult mai uşor să simulezi raţiunea unui adult cu multă educaţie decât să imiţi învăţarea obişnuită a unui bebeluş", spune ea. „Tehnica de calcul este încă cea mai bună – dacă nu singura – explicaţie ştiinţifică pe care o avem pentru modul în care un obiect fizic cum este creierul se poate purta inteligent. Dar, cel puţin deocamdată, nu avem aproape nicio idee cum este posibilă creativitatea pe care o vedem la copii".

Toată lumea a auzit despre ultimele progrese în domeniul inteligenţei artificiale şi în special în domeniul învăţării automate. De asemenea am auzit cu toţii predicţii utopice sau apocaliptice despre ce presupun aceste progrese. Au preconizat fie nemurirea, fie sfârşitul lumii, şi s-a scris mult despre ambele posibilităţi. Dar cele mai sofisticate inteligenţe artificiale sunt încă departe de a putea soluţiona problemele pe care copiii de patru ani le rezolvă cu uşurinţă. În ciuda numelui impresionant, inteligenţa artificială constă în principiu din tehnici de detectare a tiparelor statistice în seturi mari de date. Învăţarea umană înseamnă mult mai mult.

Cum putem şti atât de multe despre lumea din jurul nostru? învăţăm foarte mult chiar şi atunci când suntem bebeluşi; copiii de patru ani ştiu deja lucruri despre plante, animale şi maşini; dorinţe, credinţe şi emoţii; chiar şi despre dinozauri şi nave spaţiale.

Ştiinţa ne-a prelungit cunoştinţele despre lume până la ceea ce este inimaginabil de mare şi ceea ce este infinitezimal de mic, până la marginea universului şi începutul timpului. Şi folosim aceste cunoştinţe pentru a face clasificări şi predicţii noi, pentru a imagina posibilităţi noi şi pentru a face să se întâmple lucruri noi în lume. Dar tot ce ajunge la noi din întreaga lume este un flux de fotoni care ne loveşte retina şi unde de presiune în aer care ajung la timpanele noastre. Cum învăţăm atât de mult despre lume când dovezile pe care le avem sunt atât de limitate? Şi cum facem toate acestea cu cele câteva kilograme de mâzgă cenuşie care pluteşte în spatele ochilor noştri?

Cel mai bun răspuns de până acum este că creierele noastre efectuează calcule pe datele concrete, particulare şi dezordonate care ajung la simţurile noastre, iar aceste calcule construiesc nişte reprezentări precise ale lumii. Reprezentările par să fie structurate, abstracte şi ierarhice; includ percepţia unor obiecte tridimensionale,

gramaticile care stau la baza limbajului şi capacităţi mentale precum „teoria minţii”, care ne permit să înţelegem ce cred oamenii. Aceste reprezentări ne permit să facem o gamă largă de predicţii şi să ne imaginăm multe posibilităţi noi într-un mod creativ specific omului.

Acest tip de învăţare nu este singurul tip de inteligenţă, dar este unul destul de important pentru fiinţele umane. Şi este tipul de inteligenţă care este specialitatea copiilor cu vârste foarte fragede. Deşi copiii sunt extrem de slabi la planificarea şi luarea deciziilor, ei sunt creaturile care învaţă cel mai bine în întregul univers. O mare parte a procesului de a transforma datele în teorii se întâmplă înainte de vârsta de cinci ani.

De la Aristotel şi Platon au existat două modalităţi de bază de a aborda problema cum ştim ceea ce ştim şi acestea sunt încă principalele abordări în învăţarea automată. Aristotel a abordat problema de jos în sus: începe cu simţurile – fluxul de fotoni şi de vibraţii în aer (sau pixelii ori sunetele unei imagini sau înregistrări digitale) – şi vezi dacă poţi extrage tipare de acolo. Această abordare a fost dusă mai departe de asociaţionişti clasici, precum filosofii David Hume şi J.S. Mill, şi, mai târziu, de psihologii behaviorişti, ca Pavlov şi B.F. Skinner. În această perspectivă, caracterul abstract şi structurile ierarhice ale reprezentărilor sunt aproape iluzorii sau cel puţin un epifenomen. Toată munca poate fi făcută prin asociere şi detectarea tiparelor – în special dacă există suficiente date.

În timp, a existat o luptă între această abordare de jos în sus a misterului învăţării şi alternativa lui Platon, de sus în jos. Poate că obţinem cunoştinţe abstracte din date concrete pentru că ştim deja multe şi, în special, pentru că avem deja o multitudine de concepte abstracte de bază, mulţumită evoluţiei. Asemenea cercetătorilor, putem folosi acele concepte pentru a formula ipoteze despre lume. Apoi, în loc să încercăm să extragem tipare din datele brute, putem face predicţii despre cum ar trebui să arate datele dacă aceste ipoteze sunt corecte. Alături de Platon, au folosit această abordare şi filosofi, şi psihologi „raţionalişti”, precum Descartes şi Noam Chomsky.

Iată exemplul de zi cu zi care ilustrează diferenţa dintre cele două metode: soluţionarea potopului de poştă electronică nedorită. Datele constau într-o listă lungă şi nesortată de mesaje din inboxul tău. Realitatea este că unele dintre aceste mesaje sunt relevante, şi nu spam. Cum putem utiliza datele pentru a face diferenţa între ele?

Să luăm tehnica de jos în sus mai întâi. Observi că mesajele spam tind să aibă anumite caracteristici: o listă lungă de adrese, origini în Nigeria, referinţe la premii de milioane de dolari sau Viagra. Problema este că şi unele mesaje foarte utile pot avea aceste caracteristici. Dacă analizezi suficiente exemple de e-mailuri spam şi nonspam, poţi vedea nu doar că mesajele spam au acele caracteristici, dar şi că ele tind să

apară împreună (Nigeria plus un milion de dolari înseamnă probleme). De fapt, ar putea exista corelaţii subtile de nivel superior care diferenţiază mesajele spam de cele utile – un anumit tipar de greşeli de scriere şi adrese IP, de exemplu. Dacă detectezi aceste tipare, poţi filtra mesajele spam.

Tehnicile de învăţare automată de jos în sus fac acest lucru. „Botul" care învaţă primeşte milioane de exemple, fiecare cu un anumit set de caracteristici şi fiecare etichetat ca fiind spam (sau altă categorie) sau nu. Computerul poate extrage tiparul de caracteristici care le diferenţiază, chiar dacă este foarte subtil.

Dar dacă folosim abordarea de sus în jos? Primesc un e-mail de la editorul revistei *Journal of Clinical Biology*. Se referă la una dintre lucrările mele şi spune că vrea să publice un articol scris de mine. Fără Nigeria, fără Viagra, fără milionul de dolari; e-mailul nu are niciuna dintre caracteristicile mesajelor spam. Dar, folosind ce ştiu deja şi gândind abstract despre procesul care produce mesajele spam, pot să-mi dau seama că acest e-mail este suspect:

1. *Ştiu că cei care produc spam încearcă să fure bani de la alţii exploatând lăcomia umană.*
2. *Mai ştiu şi că revistele legitime „cu acces deschis" au început să-şi acopere costurile taxându-i pe autori, şi nu pe abonaţi, şi că nu sunt specialist în biologie clinică.*

Dacă alătur aceste informaţii, pot să produc o ipoteză nouă despre originea acelui e-mail. Este conceput pentru a-i păcăli pe autori să plătească pentru „a publica" un articol într-un jurnal fals. E-mailul a fost rezultatul aceluiaşi proces dubios pe care l-au suportat şi alte e-mailuri spam, chiar dacă nu arată deloc ca un spam obişnuit. Pot să trag această concluzie dintr-un singur exemplu şi pot continua să-mi testez ipoteza şi mai departe, dincolo de orice informaţie din corpul e-mailului, căutând pe internet numele „editorului".

În termeni IT, am început cu un „model generativ" care include concepte abstracte precum lăcomia şi înşelăciunea şi descrie procesul care produce păcăleli pe e-mail. Acest lucru îmi permite să recunosc e-mailul spam clasic cu prinţul nigerian, dar îmi permite, de asemenea, să-mi imaginez multe tipuri diferite de spamuri posibile. Când primesc e-mailul de la aşa-zisa revistă, pot să merg înapoi pe firul raţionamentului: „Acesta pare tipul de e-mail care iese dintr-un proces de generare a spamurilor".

Noul interes sporit pentru inteligenţa artificială vine datorită faptului că cercetătorii din domeniu au produs recent versiuni puternice şi eficiente ale ambelor acestor metode de învăţare. Dar nu este nimic cu adevărat nou legat de metodele însele.

Învăţarea profundă ascendentă

În anii 1980, informaţidenii au conceput o metodă ingenioasă de a face computerele să detecteze tipare în date: arhitectura conecţionistă sau reţeaua neuronală (partea „neuronală” era şi încă este metaforică). Abordarea nu a mai fost folosită în anii 1990, dar a fost reînviată recent cu metode puternice de „învăţare profundă”, cum ar fi Google Deep Mind.

De exemplu, îi poţi da unui program de învăţare profundă mai multe imagini de pe internet cu eticheta „pisică”, altele cu eticheta „casă” şi aşa mai departe. Programul poate detecta tiparele care diferenţiază cele două seturi de imagini şi folosi apoi acele informaţii pentru a eticheta corect imagini noi. Anumite tipuri de învăţare automată, denumite învăţarea nesupravegheată, pot detecta tipare în date fără niciun fel de etichetă; pur şi simplu, caută grupuri de caracteristici – ceea ce oamenii de ştiinţă numesc analiză a factorilor. În maşinile de învăţare profundă, aceste procese sunt repetate la niveluri diferite. Unele programe pot chiar să descopere caracteristici relevante din datele brute ale pixelilor sau sunetelor; computerul poate începe prin detectarea tiparelor din imaginea brută care corespund marginilor şi liniilor, apoi pot găsi tipare în acele tipare care corespund feţelor şi aşa mai departe.

Altă tehnică ascendentă cu istorie lungă este învăţarea cu întărire. În anii 1950, B.F. Skinner, bazându-se pe lucrările lui John Watson, a intrat în istorie pentru programarea unor porumbei să facă acţiuni elaborate, chiar şi să ghideze torpile aeriene către ţinte (un ecou îngrijorător al inteligenţei artificiale recente), oferindu-le un anumit program de recompense şi pedepse. Ideea principală era că acţiunile care sunt recompensate vor fi repetate, iar cele pedepsite nu, până la obţinerea comportamentului dorit. Chiar şi în zilele lui Skinner, acest proces simplu, repetat de mai multe ori, putea conduce la comportamente complexe. Computerele sunt concepute să efectueze operaţii simple în mod repetat, la o scară ce lasă în urmă imaginaţia omenească, iar sistemele de calcul pot învăţa competenţe remarcabil de complexe în acest fel.

De exemplu, cercetătorii de la Google Deep Mind au folosit o combinaţie de învăţare profundă şi învăţare cu întărire pentru a învăţa un computer să joace jocuri video Atari. Computerul nu ştia nimic despre cum funcţionează jocurile. A început prin a acţiona aleatoriu şi singurele informaţii pe care le primea erau cum arată ecranul la un moment dat şi care este scorul. Învăţarea profundă a ajutat la interpretarea caracteristicilor de pe ecran, iar învăţarea prin întărire recompensa sistemul pentru scoruri mari. Computerul a devenit foarte bun la câteva dintre jocurile jucate, dar, de asemenea, nu a putut să înveţe nimic la altele pe care oamenii reuşeau să le joace la fel de uşor

ca pe celelalte.

O combinaţie similară de învăţare profundă şi învăţare cu întărire a permis succesul pe care l-a obţinut Deep Mind cu Alpha Zero, un program care a reuşit să-i învingă pe jucătorii umani atât la şah, cât şi la Go, echipat doar cu o cunoaştere a regulilor jocului şi câteva capacităţi de planificare. Alpha Zero mai are o caracteristică interesantă: funcţionează jucând sute de milioane de jocuri cu el însuşi. Pe măsură ce face acest lucru, izolează greşelile care au condus la învingeri şi repetă şi elaborează strategiile care au condus la victorii. Astfel de sisteme, ca şi altele care implică tehnici numite reţele adversariale generative, generează date, dar şi observă date.

Când ai puterea computaţională pentru a aplica aceste tehnici la seturi de date foarte mari sau pe milioane de mesaje e-mail, imagini de pe Instagram sau înregistrări vocale, poţi rezolva probleme care păreau foarte dificile înainte. Aceasta este sursa unei mari părţi din entuziasmul care pluteşte acum în domeniul informaticii. Dar merită să ne amintim că acele probleme – cum ar fi recunoaşterea pisicii dintr-o imagine sau că un cuvânt este pronunţat de Siri – sunt triviale pentru un bebeluş uman. Una dintre cele mai interesante descoperiri din informatică este că problemele care sunt uşoare pentru noi (cum ar fi identificarea pisicilor) sunt dificile pentru computere, mult mai dificile decât a juca şah sau Go. Computerele au nevoie de milioane de exemple pentru a clasifica obiecte pe care noi le clasificăm cu doar câteva exemple. Sistemele ascendente pot generaliza până la exemple noi, pot eticheta o imagine nouă ca fiind o pisică destul de precis, în general. Dar fac acest lucru destul de diferit de modul în care îi generalizează pe oameni. Unele imagini aproape identice unei pisici nu vor fi identificate de noi ca pisici, iar altele care par a fi o pată aleatorie vor fi.

Modelele bayesiene descendente

Abordarea descendentă a jucat un rol important la începuturile inteligenţei artificiale şi, în anii 2000, a experimentat şi ea o renaştere, sub forma modelelor generative probabilistice sau bayesiene.

Încercările timpurii de a utiliza această abordare au făcut faţă la două tipuri de probleme. În primul rând, majoritatea tiparelor de dovezi pot fi explicate în principiu prin mai multe ipoteze: este posibil ca e-mailul meu referitor la jurnal să fie veritabil, doar că nu pare. În al doilea rând, care este originea conceptelor pe care le utilizează modelul generativ? Platon şi Chomsky spun că ne naştem cu ele. Dar cum putem explica modul în care învăţăm ultimele concepte din ştiinţă? Sau cum înţeleg copiii tineri conceptele de dinozaur sau navă spaţială?

Modelele bayesiene combină testarea modelelor generative şi ipotezelor cu teoria probabilităţilor şi abordează aceste două probleme. Un model bayesian îţi permite să calculezi cât de probabil este ca o anumită ipoteză să fie adevărată, date fiind anumite informaţii. Şi, făcând ajustări mici, dar sistematice la modelele pe care le avem deja şi testându-le cu datele disponibile, putem face uneori concepte şi modele noi din cele vechi. Dar aceste avantaje sunt umbrite de alte probleme. Tehnicile bayesiene te pot ajuta să alegi care ipoteză din două este mai probabilă, şi niciun sistem nu le poate lua în considerare în mod eficient pe toate. Cum decizi ce ipoteze merită testate în primul rând?

Brenden Lake de la NYU şi colegii săi au utilizat aceste metode descendente pentru a rezolva altă problemă care este uşoară pentru oameni, dar extrem de dificilă pentru computere: recunoaşterea caracterelor neclare scrise de mână. Priveşte un caracter scris pe un pergament japonez. Chiar dacă nu l-ai mai văzut niciodată, probabil poţi să-ţi dai seama dacă este similar sau diferit de un caracter de pe alt pergament japonez. Probabil poţi să-l desenezi şi poţi chiar să faci un caracter japonez fals în baza a ceea ce vezi, unul care va arăta destul de diferit de un caracter coreean sau rusesc.[60]

Metoda ascendentă pentru recunoaşterea caracterelor scrise de mână este să dai computerului mii de exemple pentru fiecare şi să-l laşi pe el să extragă caracteristicile principale. În schimb, Lake şi colaboratorii săi au dat programului un model general pentru cum poţi desena un caracter: o mişcare din condei merge în stânga sau în dreapta; după ce termini una, începi alta şi aşa mai departe. Când programul vedea un anumit caracter, putea deduce cea mai probabilă secvenţă de mişcări care a condus la el – la fel cum eu am dedus că procesul de spam a condus la e-mailul meu dubios. Apoi poate judeca dacă un caracter nou putea rezulta din acea secvenţă sau din alta şi putea produce el însuşi un set similar de mişcări. Programul funcţiona mult mai bine decât un program de învăţare profundă aplicat pe exact aceleaşi date şi oglindea îndeaproape performanţa fiinţelor umane.

Aceste două abordări ale învăţării automate au puncte forte şi puncte slabe complementare. În abordarea ascendentă, programul nu are nevoie de prea multe cunoştinţe la început, dar are nevoie de foarte multe date şi poate generaliza doar într-un mod limitat. În abordarea descendentă, programul poate învăţa din numai câteva exemple şi poate face generalizări mult mai largi şi mai variate, utilizând învăţarea profundă pentru a implementa deducţiile bayesiene.

Succesul recent al inteligenţei artificiale este în parte rezultatul extinderilor acestor idei vechi. Dar are mai mult de-a face cu faptul că, mulţumită internetului, avem mult mai multe date şi, mulţumită legii

lui Moore, avem mult mai multă putere de calcul pe care o putem aplica acestor date. În plus, subapreciem faptul că datele pe care le avem au fost deja sortate şi prelucrate de fiinţe umane. Imaginile cu pisici postate pe web sunt imagini canonice cu pisici – imagini pe care oamenii le-au ales deja ca fiind imagini „bune". Google Translate funcţionează pentru că profită de milioane de traduceri umane şi le generalizează într-o bucată nouă de text, în loc să înţeleagă cu adevărat frazele în sine.

Dar lucrul cu adevărat remarcabil despre copiii umani este că reuşesc cumva să combine cele mai bune caracteristici ale fiecărei abordări, apoi trec mult mai departe de ele. În ultimii cincisprezece ani, adepţii psihologiei dezvoltării au explorat modul în care copiii află structuri din date. Copiii de patru ani pot învăţa din unul sau două exemple, la fel cum învaţă un sistem descendent, şi pot generaliza concepte foarte diferite. Dar pot învăţa şi concepte şi modele noi de la datele în sine, aşa cum face un sistem ascendent.

De exemplu, în laboratorul nostru, dăm copiilor tineri un „detector de blicheţi[61]" – o maşină nouă ce trebuie descoperită, una pe care nu au mai văzut-o niciodată. Este o cutie care se aprinde şi redă muzică când pui anumite obiecte în ea, şi nu altele. Le dăm copiilor unul sau două exemple pentru cum funcţionează maşina, arătându-le că, de exemplu, două cuburi roşii o activează, în timp ce o combinaţie între verde şi galben nu o activează. Chiar şi copiii în vârstă de optsprezece luni pricep imediat principiul general, şi anume că cele două obiecte trebuie să fie identice pentru a activa cutia, şi generalizează acel principiu la exemple noi: de exemplu, vor alege două obiecte care au aceeaşi *formă* pentru a face maşina să funcţioneze. În alte experimente, am arătat despre copii că pot pricepe că anumite proprietăţi ascunse invizibile fac maşina să funcţioneze sau că maşina funcţionează în baza unui principiu logic abstract.[62]

Putem verifica acest principiu şi în procesul de învăţare obişnuit al copiilor. Copiii tineri învaţă rapid teorii intuitive abstracte din biologie, fizică şi psihologie în aproape acelaşi fel în care o fac şi oamenii de ştiinţă adulţi, chiar şi cu relativ puţine date.

Realizările remarcabile ale învăţării automate ale sistemelor recente de inteligenţă artificială – atât ascendente, cât şi descendente – au loc într-un spaţiu restrâns şi bine definit de ipoteze şi concepte – un set precis de componente de joc şi mutări, un set predeterminat de imagini. Comparativ, copiii şi oamenii de ştiinţă îşi schimbă uneori radical conceptele, realizând schimbări de paradigmă în loc să ajusteze pur şi simplu conceptele pe care le au deja.

Copiii de patru ani pot recunoaşte imediat pisicile şi pot înţelege cuvintele, dar pot, de asemenea, face deducţii creative şi surprinzătoare care trec mult dincolo de experienţele lor. Propriul

meu nepot mi-a explicat recent, de exemplu, că, dacă un adult vrea să devină din nou un copil, ar trebui să încerce să nu mănânce legume sănătoase, deoarece legumele sănătoase îi fac pe copii să crească şi să se transforme în adulţi. Acest tip de ipoteză, una plauzibilă la care nu s-ar gândi un adult, este caracteristic copiilor tineri. De fapt, eu şi colegii mei am arătat sistematic că preşcolarii sunt mai buni la crearea de ipoteze improbabile decât copiii mai în vârstă sau decât adulţii.[63] Nu avem aproape nicio idee cum este posibil acest tip de învăţare creativă şi inovare.

Privind ce fac copiii însă, programatorii pot face observaţii utile despre direcţiile învăţării pentru computere. Două caracteristici ale procesului de învăţare al copiilor sunt izbitoare. Copiii învaţă activ, nu absorb pasiv date, aşa cum face inteligenţa artificială. La fel cum experimentează oamenii de ştiinţă, copiii sunt motivaţi intrinsec să extragă informaţii din lumea înconjurătoare prin explorarea şi joaca lor fără de sfârşit. Studiile recente arată că această explorare este mai sistematică decât pare şi este bine adaptată pentru a găsi dovezi convingătoare pentru a suporta formarea ipotezelor şi alegerea teoriilor.[64] Integrarea curiozităţii în maşini şi permisiunea de a interacţiona activ cu lumea pot îi un drum mai realist către învăţarea diversificată.

În al doilea rând, copiii, spre deosebire de inteligenţele artificiale existente, învaţă social şi cultural. Oamenii nu învaţă în izolare, ci se folosesc de înţelepciunea acumulată a generaţiilor trecute. Studiile recente arată că până şi preşcolarii învaţă prin imitaţie şi ascultând mărturiile altora. Dar nu îşi ascultă pur şi simplu pasiv profesorii. În schimb, interiorizează informaţii de la alte persoane într-un mod remarcabil de subtil şi sensibil, făcând deducţii complexe despre locul din care vine informaţia şi cât este de fiabilă, integrându-şi sistematic propriile experienţe în ceea ce aud.[65]

„Inteligenţa artificială” şi „învăţarea automată” sună înspăimântător. Şi, în anumite feluri, sunt. Aceste sisteme sunt utilizate pentru a controla arme, de exemplu, şi de asta chiar ar trebui să ne temem. Totuşi, prostia naturală poate cauza dezastre mult mai mari decât inteligenţa artificială; noi, oamenii, va trebui să fim mult mai inteligenţi decât am fost în trecut pentru a gestiona corespunzător noile tehnologii. Dar nu există baze solide nici pentru viziunea apocaliptică, nici pentru cea utopică, în care inteligenţa artificială îi înlocuieşte pe oameni. Cât timp nu vom rezolva paradoxul elementar al învăţării, cele mai bune inteligenţe artificiale nu vor putea concura cu un copil obişnuit de patru ani.

Capitolul 22

Peter Galison

Peter Galison este istoric al ştiinţei; profesor universitar Joseph Pellegrino şi cofondator al iniţiativei Black Hole *de la Universitatea Harvard; şi autorul cărţii* Einsteins Clocks Poincaré's Maps: Empires of Time.

În calitate de istoric al ştiinţei, Peter Galison se concentrează, în general, pe intersecţia teoriei cu experimentul.

„De ani buni, îmi ghidez activitatea după confruntarea ciudată a ideilor abstracte cu obiecte extrem de concrete", mi-a spus el cândva, explicând care este părerea lui despre ceea ce face. La întâlnirea din Washington, Connecticut, a discutat despre tensiunile din Războiul Rece dintre ingineri (ca Wiener) şi administratorii proiectului Manhattan (ca Oppenheimer): „Când Wiener ne avertizează despre pericolele ciberneticii, pe de o parte, încearcă să concureze cu limbajul neobişnuit pe care îl folosesc oamenii ca Oppenheimer: «Când am văzut explozia de la Trinity, m-am gândit la Bhagavad Gītā – Eu sunt moartea, distrugătoarea lumilor». Acest sens, şi anume că fizica poate vorbi despre natura universului şi politicile militare aeriene, era respingător şi seducător. Într-un fel, se poate vedea acest lucru în mod repetat în ultimele decenii – nanoştiinţele, ADN recombinant, cibernetica: «Vă raportez despre ştiinţa care aduce cu ea promisiunea salvării şi pericolul anihilării – şi trebuie să fiţi atenţi, pentru că vă poate omorî». Este o poveste foarte seducătoare şi este repetată în inteligenţa artificială şi în robotică".

Când aveam douăzeci şi patru de ani şi am intrat prima dată în contact cu ideile lui Wiener, apoi i-am cunoscut colegii la întâlnirile de la MIT pe care le descriu în introducerea cărţii, eram foarte puţin interesat de avertismentele sau admonestările lui Wiener. Curiozitatea mea era împinsă de la spate de natura puternică şi radicală a perspectivei sale asupra vieţii, bazate pe teoria matematică a comunicaţiilor în care mesajul este neliniar: după Wiener, „conceptele noi ale comunicării şi controlului au implicat o nouă interpretare a omului, a cunoştinţelor omului despre univers şi a societăţii". Şi acest lucru a condus la prima mea carte, care a luat teoria informaţiei – teoria matematică a comunicării – ca model pentru toată experienţa umană.

Într-o conversaţie recentă, Peter mi-a spus că începe să scrie o carte – despre construire, prăbuşire şi gândire – care analizează natura de cutie neagră a ciberneticii şi cum reprezintă aceasta ceea ce percepe ca „transformarea fundamentală a învăţării,

învăţării automate, ciberneticii şi sinelui”.

În a doua cea mai bună carte a sa, marele matematician medieval al-Khwarizmi a descris noua formă de aritmetică de origine indiană. Numele său, în scurt timp corelat cu „algorismus” (în latina medievală târzie), a ajuns să desemneze procedurile aplicate numerelor, alunecând treptat către „algoritm” (pe modelul „logaritm”), în franceză şi apoi în engleză. Dar îmi place ideea de algoritmician modern, chiar dacă această idee nu-i place şi corectorului meu gramatical. Prin acest termen înţeleg o persoană profund suspicioasă cu privire la intervenţia judecăţii umane, cineva care foloseşte această judecată pentru a încălca normele fundamentale a ceea ce înseamnă să fii obiectiv (şi, prin urmare, ştiinţific).

Aproape de sfârşitul secolului XX, o lucrare scrisă de doi psihologi de la Universitatea din Minnesota a rezumat o literatură vastă care a tulburat mult timp apele prognozelor. O parte, spuneau ei, a ţinut prea mult timp cu dinţii – şi, în cele din urmă, fără etică – de „metoda clinică” de prognoză, care punea accent pe tot ce este subiectiv: „informai”, „în cap” şi „impresionist”. Aceşti clinicieni erau oameni (spuneau psihologii) care credeau că pot să-şi studieze subiecţii cu o grijă meticuloasă, se pot aduna în comisii şi pot face predicţii pe bază de judecată cu privire la recidivismul criminal, succesul la facultate, rezultatele medicale şi altele asemenea. Cealaltă parte, au continuat psihologii, întrupa tot ce nu întrupau clinicienii, îmbrăţişând obiectivul: „formal”, „mecanic”, „algoritmic”. Autorii au considerat că acest lucru stă la rădăcina întregului triumf al ştiinţei de după Galilei. Nu doar că ştiinţa beneficia de pe urma statisticii, dar şi, în mare măsură, ştiinţa era statisticianul mecanic. Răsfoind 136 de studii pe tema predicţiilor, din domenii diverse de la sentinţele judecătoreşti la psihiatrie, autorii au arătat că, în 128 dintre ele, predicţiile ce conţineau tabele statistice, o ecuaţie de regresie multiplă sau o judecată algoritmică au egalat sau depăşit în acurateţe pe cele care au folosit abordarea subiectivă.

Au continuat prin catalogarea a şaptesprezece justificări eronate pentru susţinerea abordării clinice. Aceştia erau trântorii care-şi vedeau propriile interese şi care se temeau că maşinile îi vor lăsa fără locurile lor comode de muncă. Alţii nu aveau educaţia necesară să urmărească argumentele statistice. Un grup nu avea încredere în formalizarea matematicii; alţii au criticat sever ceea ce ei considerau că este o „dezumanizare statistică”; iar alţii au spus că scopul este înţelegerea, nu predicţia. Dar, oricare ar fi motivaţiile, rezumatul a conchis că este de-a dreptul imoral să reţii puterea obiectivităţii asupra subiectivităţii, a algoritmilor asupra judecăţii experţilor.[66]/

Perspectiva algoritmicianului a câştigat teren. Anne Milgram a fost procuror general al statului New Jersey din 2007 până în 2010.

Când a preluat postul, a vrut să ştie pe cine aresta, acuza şi întemniţa statul şi pentru ce infracţiuni. La momentul respectiv, raportează ea într-o conferinţă TED ulterioară, datele sau statisticile erau aproape inexistente. Impunând predicţiile statistice, continuă ea, forţele de ordine din Camden în timpul mandatului său au reuşit să reducă rata crimelor comise cu 41%, salvând 37 de vieţi, reducând rata totală a crimelor la 26 de procente. După ce s-a alăturat Fundaţiei Arnold în calitate de vicepreşedinte pentru justiţie penală, a stabilit o echipă de administratori de date şi statisticieni pentru a crea un instrument de evaluare a riscurilor; de fapt, a interpretat misiunea echipei ca fiind aceea de a decide cum vor fi întemniţate „persoanele periculoase" şi cum vor fi eliberate cele nepericuloase. „Motivul care stătea în spatele acestei decizii", a continuat Milgram, „este modul în care luăm deciziile. Judecătorii au cele mai bune intenţii când iau aceste decizii cu privire la riscuri, dar le iau subiectiv. Erau ca evaluatorii de talente din baseball de acum douăzeci de ani care încercau să îşi folosească instinctul şi experienţa pentru a încerca să stabilească ce riscuri prezintă cineva. Erau subiectivi şi ştim ce se întâmplă când oamenii iau decizii subiective, anume că ne înşelăm adesea". Echipa sa a stabilit peste nouă sute de factori de risc, dintre care nouă erau nişte predictori foarte buni. Cele mai urgente întrebări pentru echipă erau: o persoană va comite o nouă infracţiune? Persoana respectivă va comite un act de violenţă? Cineva va reveni în faţa instanţei? Avem nevoie, a concluzionat Milgram, de o „măsură obiectivă pentru risc" care ar trebui aplicată de judecata judecătorilor. Ştim că procesul statistic algoritmic funcţionează. Acesta este motivul pentru care „Google este Google", spune ea, şi motivul pentru care vânătoarea de talente câştigă meciurile.[67]/

Algoritmicienii au câştigat. Ne-am obişnuit cu ideea că protocoalele şi datele pot şi trebuie să ne ghideze în acţiunile noastre de zi cu zi, de la mementouri despre locurile în care trebuie să ajungem până la probabilitatea de a fi victima unei infracţiuni. Până acum, conform literaturii, dimensiunile legală, etică, oficială şi economică ale algoritmilor sunt toate cvasi-infinite. Aş vrea să mă concentrez mai mult pe un cântec de sirenă al algoritmilor: promisiunea lor de obiectivitate.

Obiectivitatea ştiinţifică are o istorie. Acest lucru poate părea surprinzător. Nu este corectă noţiunea exprimată mai sus de psihologii din Minnesota? Nu este obiectivitatea coextensivă cu ştiinţa însăşi? Aici merită să facem un pas înapoi pentru a reflecta asupra tuturor virtuţilor epistemice cărora le-am putea acorda valoare în activitatea ştiinţifică. Calificarea pare ceva bun de avut; la fel şi predicţia, explicaţia, unificarea, precizia, acurateţea, siguranţa şi utilitatea pedagogică. În cea mai bună dintre lumile posibile, aceste virtuţi

epistemice ar trage toate în aceeaşi direcţie. Dar nu fac asta – nu mai mult decât coincid în mod necesar virtuţile noastre etice. Recompensarea oamenilor în funcţie de nevoile lor poate intra în conflict cu recompensarea oamenilor în funcţie de competenţe. Egalitate, echitate şi meritocraţie – etica, într-un anumit sens, este legată de adjudecarea bunurilor conflictuale. Prea des uităm că acest conflict există şi în ştiinţă. Proiectează un instrument pentru a fi cât se poate de sensibil şi adesea acesta va fluctua mult, făcând imposibilă repetarea unei măsurări.

„Obiectivitatea ştiinţifică" a pătruns atât în practica, precum şi în nomenclatura ştiinţifică după prima treime a secolului al XIX-lea. Se poate vedea clar în atlasele ştiinţifice care au pus la baza oamenilor de ştiinţă obiectele de bază ale specialităţii lor: au existat şi există atlase ale mâinii, ale craniului, ale norilor, cristale, flori, tablouri de cameră, emulsii nucleare şi boli ale ochilor. În secolul al XVIII-lea, era evident că nu vom înfăţişă o anumită frunză de trifoi, arsă de soare şi mestecată de omizi, pe care o găsim în faţa casei, într-un atlas. Nu, ţinta era – dacă erai un filosof natural genial precum Goethe, Albinius sau Cheselden – să observăm natura, dar apoi să perfecţionăm obiectul observat, să-l abstractizăm vizual până la ideal. Să luăm, de exemplu, un schelet, să-l vizualizăm printr-o cameră lucidă şi să-l desenăm cu atenţie, apoi să corectăm „imperfecţiunile". Avantajul acestei despărţiri a cortinelor de simplă experienţă era clar: punea la dispoziţie un ghid universal, unul neataşat de inconsecvenţele unei variaţii individuale.

Pe măsură ce volumul ştiinţei a crescut, şi numărul oamenilor de ştiinţă la fel, partea proastă a idealizării a devenit mai clară. Era ceva când Goethe înfăţişă „planta primordială" sau „insecta primordială". Dar era cu totul altceva să avem miriade de cercetători diferiţi care-şi fixează imaginile în moduri diferite şi uneori contradictorii. Gradual, începând cu anii 1830, începem să observăm ceva nou: pretenţia că realizarea imaginii era făcută cu o intervenţie umană minimă, că erau respectate protocoale. Acest lucru putea însemna conturarea unei frunze cu creionul sau apăsarea acesteia în tuş şi transferarea ei pe pagină. Mai însemna şi că ne mândream cu înfăţişarea imaginii prin microscop a unui obiect natural chiar şi cu imperfecţiunile sale. Aceasta era o idee radicală: fulgi de zăpadă înfăţişaţi fără o simetrie hexagonală perfectă, distorsiunea culorii pe marginea unei lentile de microscop, ţesut rupt în jurul marginilor în procesul pregătirii sale.

Obiectivitatea ştiinţifică a ajuns să însemne că reprezentările noastre erau executate prin eliminarea intervenţiei – chiar dacă asta însemna să reproducem culoarea galbenă de lângă marginea imaginii sub microscop, în ciuda faptului că omul de ştiinţă ştia că decolorarea era cauzată de lentilă, şi nu o caracteristică a obiectului cercetat.

Avantajul obiectivităţii era clar: înlocuia dorinţa de a vedea o teorie realizată sau o perspectivă acceptată confirmată. Dar obiectivitatea avea costurile sale. Pierdeam reprezentarea artistică precisă, care se putea învăţa uşor, colorată şi care cuprindea întreaga profunzime a câmpului în cazul unui cadavru disecat. Obţineam o fotografie înceţoşată, cu o profunzime slabă a câmpului, alb-negru, pe care niciun student la medicină (şi nici mulţi colegi din medicină) nu o puteau folosi pentru a învăţa şi compara cazuri. Cu toate acestea, o lungă parte din secolul al XIX-lea, virtutea obiectivităţii cu intervenţie redusă era în creştere.

Începând cu anii 1930, obiectivitatea ştiinţifică strictă din reprezentarea ştiinţifică a început să aibă probleme. În catalogarea spectrelor stelare, de exemplu, niciun algoritm nu putea concura cu observatorii antrenaţi, care le puteau sorta cu o acurateţe şi replicabilitate mult mai mari decât orice procedură care urmează reguli. La sfârşitul anilor 1940, doctorii au început să înveţe să citească electroencefalogramele. Era necesară o părere de expert pentru sortarea diverselor tipuri de citiri ale atacurilor cerebrale, în timp ce niciuna dintre încercările dinainte de a utiliza analiza frecvenţelor nu se potrivea cu judecata lor. Magnetogramele solare – cartografierea câmpurilor magnetice de pe Soare – necesita din partea expertului să diferenţieze semnalul real de artefactele care luau naştere din cauza instrumentelor de măsurare. Chiar şi fizicienii care se ocupă de particule au recunoscut că nu puteau programa un computer pentru a sorta anumite tipuri de urme în categoriile lor speciale; era necesară o părere de profesionist.

Nu trebuie să existe confuzie aici: aceasta nu era o întoarcere la geniul invocat al unui idealist din secolul al XVIII-lea. Nimeni nu credea că te puteai antrena să devii un Goethe care singur printre alţi cercetători putea alege forma universală şi ideală a unei plante, a unei insecte sau a unui nor. Expertiza se poate deprinde – poţi face un curs pentru a învăţa să faci judecăţi de expert în ceea ce priveşte electroencefalogramele, spectrele stelare sau urmele din camerele cu bule; din păcate, nimeni nu s-a gândit că ai putea absolvi un curs care te-ar putea face un observator excepţional. Nu există niciun drum regal către a deveni un Goethe. În nenumărate adase ştiinţifice se pot vedea argumente explicite că factorii „subiectivi” trebuie să facă parte din activitatea ştiinţifică necesară pentru a crea, clasifica şi interpreta imaginile ştiinţifice.

Ce vedem în atât de multe pretenţii ale algoritmicienilor este o dorinţă uriaşă de a găsi obiectivitatea ştiinţifică fix prin abandonarea judecăţii şi bizuirea pe proceduri mecanice – în numele obiectivităţii ştiinţifice. Multe state americane au legiferat utilizarea unor algoritmi de stabilire a sentinţelor şi cauţiunilor. Mai bine o maşină, se

argumentează, decât inconsecvenţele judecăţii unui judecător.

Iată deci o avertizare din partea ştiinţelor. Proceduralismul algoritmic fără intervenţie umană a avut într-adevăr perioada de înflorire în secolul al XIX-lea şi, bineînţeles, încă joacă un rol în multe dintre încercările tehnice şi ştiinţifice de succes. Dar ideea că obiectivitatea mecanică, interpretată drept control strict de sine, urmează o curbă simplă şi monotonă, crescând de la clinicianul impresionist şi slab la o statistică externalizată bună, nu oferă răspuns la istoria mai interesantă şi nuanţată a ştiinţelor.

Există o lecţie şi mai importantă pe care o învăţăm de la ştiinţe. Obiectivitatea mecanică este o virtute ştiinţifică printre altele, iar ştiinţele naturale învaţă această lecţie destul de des. Trebuie să facem la fel şi în domeniile ştiinţelor legale şi sociale. Ce se întâmplă, de exemplu, când algoritmul secret şi brevetat trimite o persoană la închisoare timp de zece ani şi alta timp de cinci ani pentru aceeaşi infracţiune? Rebecca Wexler, profesor consultant la proiectul societăţii informatice de la Facultatea de Drept Yale, a explorat această întrebare şi costul uriaş pe care algoritmii secretelor comerciale îl impun pentru posibilitatea de apărare legală echitabilă. Într-adevăr, dintr-o varietate de motive, organele legale nu vor fi dispuse să împărtăşească algoritmii utilizaţi pentru producţia de ADN, substanţe chimice, identificarea amprentelor, ceea ce pune apărarea într-o poziţie mult slăbită pentru a-şi susţine pledoaria. În sala de judecată, obiectivitatea, secretele comerciale şi transparenţa legală pot trage în direcţii opuse. Îmi aduce aminte de un moment din istoria fizicii. Imediat după cel de-al Doilea Război Mondial, giganţii în producţia de film Kodak şi Ilford au perfecţionat un film care putea fi utilizat pentru a descoperi interacţiunile şi descompunerile particulelor elementare. Fizicienii erau încântaţi, bineînţeles – până când companiile de film le-au spus despre compoziţia filmului că este un secret comercial, astfel încât cercetătorii nu au putut avea niciodată încredere deplină că înţeleg corect procesele pe care le studiază. Dacă dotăm lucrurile cu nişte cutii negre care nu pot fi deschise, oamenii de ştiinţă vor juca un joc periculos, cu atât mai mult justiţia penală.

Alţi critici au evidenţiat cât de periculos este să ne bizuim pe formulările unei persoane acuzate (sau condamnate) sau pe alte variabile care pot deveni cu uşurinţă, în interiorul cutiei negre a stabilirii sentinţei pe bază de algoritmi, o împuternicire pentru rasă. Prin urmele experienţei noastre de zi cu zi, ne-am obişnuit cu faptul că securitatea de la aeroport este diferită pentru copiii sub doisprezece ani şi pentru adulţii peste şaptezeci şi cinci de ani. Ce factori dorim să aibă algoritmicienii în procedurile lor adesea ascunse? Educaţie? Venit? Istoricul locurilor de muncă? Ce a citit, privit, vizitat sau cumpărat o persoană? Contactele anterioare cu legea? Cum dorim să

pondereze algoritmicienii aceşti factori? Funcţiile analitice predictive predicate pe obiectivitatea mecanică au preţul lor. Uneori poate fi un preţ ce merită plătit; alteori acel preţ ar fi devastator pentru societatea echitabilă pe care dorim să o avem.

În general, pe măsură ce convergenţa algoritmilor şi a Big Data ne controlează o parte din ce în ce mai mare a vieţii, merită să ne amintim aceste două lecţii din istoria ştiinţelor: judecata nu este coaja decorticată a unei obiectivităţi acum pure a controlului de sine. Şi obiectivitatea mecanică este o virtute care concurează y cu altele, şi nu chintesenţa unei întreprinderi ştiinţifice. Acestea sunt lecţii de ţinut minte, chiar dacă algoritmicienii visează la obiectivitate.

Capitolul 23
Drepturile maşinilor

George M. Church

George M. Church este profesor de genetică Robert Winthrop la Facultatea de Medicină de la Harvard, profesor de ştiinţe ale sănătăţii şi tehnologie la MIT Harvard şi coautor (împreună cu Ed Regis) al cărţii Regenesis: How Synthetic Biology Will Reivent Nature and Ourselves.

În ultimul deceniu, ingineria genetică a prins din urmă informatica cu privire la modul în care noile iniţiative ştiinţifice ne modelează vieţile. Inginerul genetician George M. Church, un pionier al revoluţiei de citire şi scriere a biologiei, este un jucător important în acest peisaj nou de idei. Priveşte corpul ca pe un sistem de operare, inginerii luând locul biologilor tradiţionali în reutilarea componentelor învechite ale organismelor (de la atomi la organe), aproape în acelaşi fel ca la sfârşitul anilor 1970, când inginerii lucrau pentru a crea primul computer personal, asamblând plăci de circuite, unităţi de disc, monitoare etc. George a creat şi este administratorul proiectului Personal Genome, care conduce singura sursă de informaţii cu acces deschis la datele genomului uman, despre mediu şi trăsături (GET), revigorând industria străveche, acum în creştere, a ADN-ului.

Activitatea sa a fost foarte importantă în punerea bazelor iniţiativei BRAIN (Brain Research through Advancing Inovative Neurotechnologies – Cercetarea creierului prin neurotehnologii inovatoare avansate) din 2013 a preşedintelui Obama – în ajutorul îmbunătăţirii creierelor omeneşti până la punctul în care, pentru o mare parte din ceea ce avem nevoie, să nu mai avem nevoie de inteligenţe artificiale (potenţial periculoase). „Este posibil ca unele dintre proiectele Iniţiativei BRAIN să ne permită să construim

creiere omeneşti care se potrivesc mai bine cu etica noastră şi sunt capabile să execute sarcini avansate, asemenea inteligenţei artificiale", spunea George. „Cea mai sigură cale de departe este ca oamenii să reuşească să facă toate sarcinile pe care ar vrea să le delege maşinilor, dar nu suntem încă stabili pe această cale supersigură".

Mai recent, utilizarea sa de pionierat de o importanţă crucială a enzimei CRISPR (precum şi a metodelor mai bune decât CRISPR) pentru a edita genele celulelor umane este uneori omisă de media, atunci când se spune povestea originilor CRISPR.

Atitudinea lui George la adresa viitoarelor forme ale inteligenţei artificiale generale este prietenoasă, după cum se vede în eseul care urmează. În acelaşi timp, nu pierde niciodată din vedere problema siguranţei inteligenţei artificiale. Pe această temă, observa recent: „Principalul risc în domeniul inteligenţelor artificiale, din punctul meu de vedere, nu este atât de mult dacă putem înţelege matematic ce gândesc, ci dacă suntem capabili să le învăţăm comportamentul etic. Abia dacă putem să-i învăţăm pe semenii noştri comportamentul etic".

În 1950, cartea lui Norbert Wiener *The Human Use of Human Beings* era un apogeu de viziune şi speculaţie spunând că:

Maşina, asemenea djinnului, care poate învăţa şi lua decizii în baza a ceea ce a învăţat, nu va fi obligată în niciun fel să ia decizii pe care noi ar fi trebuit să le luăm sau care să fie acceptabile pentru noi… Indiferent dacă ne încredinţăm deciziile noastre maşinilor de metal sau maşinilor din carne şi oase, care sunt birouri, laboratoare vaste, armate şi corporaţii… este târziu, iar alegerea dintre bine şi rău bate la uşă.

Dar acesta era deznodământul cărţii sale şi ne-a lăsat în aşteptare de şaizeci şi opt de ani, lipsindu-i nu numai prescripţiile şi proscripţiile, dar chiar şi o „formulare a problemei" bine articulată. De atunci am văzut avertismente similare cu privire la ameninţarea maşinilor noastre, chiar şi sub forma strigătului către mase, prin filme precum *Colossus*: *The Forbin Project* (1970), *The Terminator* (1984), *The Matrix* (1999) şi *Ex Machina* (2015). Dar acum este vremea unei actualizări majore, cu perspective noi şi proaspete – notabil concentrate pe generalizări ale drepturilor „omului" şi ale nevoilor noastre existenţiale.

Îngrijorările au avut tendinţa să se concentreze pe „noi contra lor [roboţilor]" sau „mâzga gri [nanotehnologia]" sau

„Monoculturile de clone [bio]". Pentru a extrapola tendinţele curente: ce-ar fi dacă am putea face sau creşte aproape orice şi am putea proiecta orice nivel de siguranţă şi eficacitate am dori? Orice

fiinţă gânditoare (făcută din orice aranjament de atomi) poate avea acces la orice tehnologie.

Probabil ar trebui să ne concentrăm mai puţin pe noi-contra-lor şi mai mult pe drepturile fiinţelor simţitoare în faţa emergenţei unei diversităţi de minţi fără precedent. Ar trebui să exploatăm această diversitate pentru a minimiza riscurile existenţiale globale, cum ar fi supervulcanii şi asteroizii.

Dar, ar trebui oare să spunem „ar trebui"? (Precizări necesare: în acest caz şi în multe altele, când un tehnolog descrie o cale a societăţii care „se poate", „se va" sau „ar trebui să „se întâmple, acest lucru nu înseamnă neapărat că acestea sunt şi preferinţele autorului. Poate reflecta avertizarea, nesiguranţa şi/sau evaluarea detaşată.) Roboticianul Gianmarco Veruggio şi alţii au ridicat probleme de roboetică încă din 2002; Departamentul de Comerţ şi Industrie din Marea Britanie şi Institutul pentru Viitor, o subcompanie a RÂND, au ridicat probleme legate de drepturile roboţilor încă din 2006.

„Este versus ar trebui".

Este de bun-simţ să spui că ştiinţa se preocupă de „este", şi nu de „ar trebui". Perspectiva „magisteriei nesuprapunerii" a lui Stephen Jay Gould argumentează că faptele trebuie să fie complet diferite de valori. Similar, documentul *Ştiinţă şi creaţionism* din 1999, din U.S. National Academy of Sciences, observa că „ştiinţa şi religia ocupă două tărâmuri separate". Această împărţire a fost criticată de biologul evoluţionist Richard Dawkins, de mine şi de alţii. Îl putem discuta pe „ar trebui" dacă este încadrat ca „ar trebui să facem X pentru a putea obţine Y". Ce Y ar trebui să aibă o prioritate mai mare nu este ales neapărat prin vot democratic, dar poate fi ales prin vot darwinian. Sistemele de valori şi suişurile şi coborâşurile religiei diversifică, deviază şi converg la fel ca speciile vii: sub imperiul selecţiei. „Valoarea" finală (acel „ar trebui") este supravieţuirea genelor şi a memelor.

Puţine religii spun că nu există nicio conexiune între fiinţa noastră fizică şi lumea spirituală. Miracolele sunt documentate. Conflictele dintre doctrina bisericească şi Galileo sau Darwin sunt rezolvate în cele din urmă. Credinţa şi etica se găsesc din abundenţă din specia noastră şi pot fi studiate folosind metode ştiinţifice, printre care, dar nu exclusiv, FRMN, medicamente psihoactive, chestionare etc.

Foarte practic, avem de găsit regulile etice care trebuie construite, învăţate sau alese probabilistic pentru maşini din ce în ce mai inteligente şi diverse. Avem o întreagă serie de probleme ale tramvaiului. La ce număr de oameni care-şi aşteaptă moartea ar trebui computerul să decidă să schimbe macazul la o persoană? În cele din urmă, aceasta poate fi o problemă de învăţare profundă – una în care

baze de date uriaşe de fapte şi contingenţe pot fi luate în considerare, unele aparent departe de etica discutată.

De exemplu, computerul poate deduce că persoana care scapă de moarte dacă tramvaiul este lăsat să-şi urmeze cursul este un recidivist terorist condamnat, plin de patogeni care pot distruge lumea, sau un POTUS sfânt – sau parte dintr-un lanţ de evenimente mult mai elaborat în realităţi alternative detaliate. Dacă una dintre aceste descrieri al problemei pare paradoxală sau ilogică, este posibil ca autorii Problemei Tramvaiului să fi ajustat greutăţile în fiecare parte astfel încât deciziile ezitante să fie inevitabile.

Alternativ, se poate folosi redirecţionarea atenţiei pentru a măslui sistemul, astfel încât modurile de eroare să nu fie la nivelul atenţiei. De exemplu, în Problema Tramvaiului, adevărata decizie etică a fost făcută cu câţiva ani înainte, când pietonii au primit acces pe şine sau chiar înainte de asta, când am votat să cheltuim mai mult pe divertisment decât pe siguranţa publică. Întrebări care par la început străine şi tulburătoare, cum ar fi „Cine deţine noile minţi şi cine plăteşte pentru greşelile lor?”, sunt similare legilor bine stabilite cu privire la cine deţine şi plăteşte pentru greşelile unei corporaţii.

Pantele alunecoase

Putem (supra) simplifica etica pretinzând că anumite scenarii nu se vor întâmpla. Provocările tehnice sau liniile roşii aprinse care nu pot fi trecute sunt reconfortante, dar realitatea este că, odată ce beneficiile depăşesc riscurile (chiar şi puţin sau pentru scurt timp), liniile roşii se mută. Chiar înainte de naşterea lui Louise Brown, în 1978, mulţi oameni îşi făceau griji că „va deveni un mic monstru, cumva, deformată, că va fi ceva în neregulă cu ea”[68] Puţini mai văd fertilizarea in vitro aşa în prezent.

Ce tehnologii lubrifiază panta către conştiinţa compusă? Nu este vorba doar despre algoritmii de învăţare automată profundă cu Big Iron. Am reuşit să facem rozătoarele să fie mult mai bune la o diversitate de sarcini cognitive, precum şi la manifestarea altor trăsături relevante, cum ar fi persistenţa şi anxietatea scăzută. Va fi adevărat, oare, şi pentru animalele care se apropie de inteligenţa umană? Multe arată recunoaşterea sinelui în testul oglinzii – cimpanzei, bonobo, urangutani, anumiţi delfini şi anumite balene, precum şi coţofenele.

Chiar şi linia roşie aprinsă pentru manipularea oamenilor de către oameni arată multe semne de mişcare sau de rupere cu totul. Peste 2 300 de studii clinice aprobate pentru terapia cu gene sunt în desfăşurare în întreaga lume. Un obiectiv medical major este tratarea sau împiedicarea scăderii capacităţilor cognitive, în special în lumina demografiei globale care arată o tendinţă rapidă de îmbătrânire. Unele

tratamente pentru scăderea capacităţilor cognitive vor include stimulatori cognitivi (medicamente, gene, celule, transplanturi, implanturi şi altele). Acestea vor fi utilizate pentru alte indicaţii terapeutice decât cele indicate de prospecte.

Regulile competiţiilor atletice (de exemplu, interzicerea augmentării cu steroizi sau eritropoietină) nu se aplică, aşadar, competiţiilor intelectuale din lumea reală. Toate progresele realizate în domeniul scăderii capacităţilor cognitive sunt în joc pentru indicaţiile neaprobate.

Altă frontieră pentru folosirea oamenilor de către oameni este reprezentată de „organoidele cerebrale". Acum putem accelera biologia dezvoltării. Procese care în mod normal durau luni de zile se pot realiza acum în patru zile în laborator folosind reţetele potrivite de factori de transcripţie. Putem face creiere ce recapitulează diferenţele dintre oamenii născuţi cu capacităţi cognitive aberante (de exemplu, microcefalie) cu o fidelitate din ce în ce mai mare. Vascularizarea corespunzătoare (vene, artere şi capilare) ce lipsea din succesele anterioare este adăugată acum, permiţând organoidelor cerebrale să depăşească limita anterioară de submicrolitru, cu probabilitatea de a depăşi mărimea de 1, 2 litri a creierelor umane moderne (sau chiar a creierelor de elefant, de cinci litri, sau de balenă, de 8 litri).

Computere convenţionale versus hibrizi bioelectrici

Pe măsură ce ne apropiem de un nou hop în miniaturizarea din Legea lui Moore (care, cu siguranţă, nu este un zid de netrecut), vedem limitele stocasticii atomilor dopanţi în plăci de siliciu şi limitele metodelor de fabricare a grinzilor la mărimi de aproximativ zece nanometri. Problemele legate de consumul de energie sunt şi ele clare: marele Watson, câştigătorul emisiunii *Jeopardy/*, a folosit 85.000 de waţi în timp real, în vreme ce creierele omeneşti foloseau 20 de waţi fiecare. Să fim cinstiţi, corpul omenesc are nevoie de 100 de waţi pentru a funcţiona şi de 20 de ani pentru a fi construit, deci aproximativ 6 trilioane de jouli de energie pentru a „produce" un creier uman matur. Costul de producţie al unui sistem de calcul Watson este similar. Prin urmare, de ce nu înlocuiesc oamenii computerele?

În primul rând, creierele concurenţilor de la *Jeopardy*! făceau mult mai mult decât să colecteze informaţii – multe dintre acestea ar fi considerate doar distrageri de Watson (de exemplu, controlul cerebral al zâmbitului). Alte părţi însemnau o ieşire din cutie cu o transcendenţă de neconceput pentru Watson, aşa cum vedem în cele cinci lucrări ale lui Einstein din *annus mirabilis* 1905. De asemenea,

oamenii consumă mai multă energie decât minimul (100 de waţi) necesar pentru viaţă şi reproducere. Oamenii din India consumă în medie 700 de waţi de persoană; în Statele Unite, această medie este de 10.000. Ambele medii sunt sub consumul de 85.000 de waţi al lui Watson. Computerele pot începe să semene cu noi prin tehnica de calcul neuromorfică, posibil de 1.000 de ori mai bine. Dar şi creierele omeneşti pot deveni mai eficiente. Organoidul creier în sticlă se poate apropia de limita de 20 de waţi. Avantajele bazate pe idiosincrazie ale computerelor pentru matematică, stocare şi căutare, facultăţi foarte puţin utilizate de strămoşii noştri, pot fi proiectate şi reevaluate în laboratoare.

Facebook, Agenţia Naţională de Securitate şi alţii construiesc centre de stocare a informaţiilor de ordinul exabyţilor, care consumă mai mult de un megawatt şi se întind pe patru hectare, în timp ce ADN-ul poate stoca acelaşi volum de informaţii într-un miligram. În mod clar, ADN-ul nu este o tehnologie de stocare matură, dar cu Microsoft şi Tehnicolor pariind pe el, ar fi înţelept să-i acordăm atenţie. Motivul principal pentru cele şase trilioane de jouli de energie necesare pentru obţinerea unei minţi omeneşti productive sunt cei douăzeci de ani necesari instruirii.

Chiar dacă un supercomputer îşi poate „instrui" o clonă a sa în câteva secunde, costul energetic de a produce o clonă matură de siliciu este comparabil. Proiectarea unor genii din specia Homo Sapiens poate avea un mic impact asupra acestui proces lent, dar grăbirea dezvoltării şi implantarea unei memorii extinse (ca exabyţi ADN sau prin alte mijloace) poate înlocui timpul de replicare al unui computer biologic până aproape de viteza de dublare a celulelor (între unsprezece minute şi douăzeci şi patru de ore). Ideea este că, deşi noi nu ştim ce raport de bio/homo-/ nano-/robohibrizi va fi dominant în fiecare etapă a evoluţiei noastre accelerate, putem ţinti către un comportament cu grad ridicat de umanitate, corectitudine şi tratament („folosire") sigur (a) unul faţă de celălalt.

Declaraţiile de drepturi datează încă din 1689 în Anglia. Franklin Delano Roosevelt a proclamat „Cele Patru Libertăţi" – libertatea de exprimare, libertatea conştiinţei, libertatea faţă de teamă şi libertatea faţă de nevoi. Declaraţia Universală a Drepturilor Omului, emisă de Organizaţia Naţiunilor Unite în 1948, includea dreptul la viaţă; interzicerea sclaviei; apărarea drepturilor atunci când sunt încălcate; libera circulaţie; libertatea de asociere, gândire, conştiinţă şi religie; drepturi culturale, sociale şi economice; îndatoririle persoanei faţă de societate; şi interzicerea utilizării drepturilor în contradicţie cu scopurile şi principiile Naţiunilor Unite.

Natura „universală" a acestor drepturi nu este universal îmbrăţişată şi este supusă unor critici şi neconformităţi extinse. Cum

afectează emergenţa inteligenţelor non-Homo această discuţie? Cel puţin devine din ce în ce mai dificil să te ascunzi în spatele unor intuiţii vagi în ceea ce priveşte deciziile etice – „o să-mi dau seama când voi vedea” (spunea Potter Stewart, judecător la Curtea Supremă de Justiţie a Statelor Unite în 1964) sau „înţelepciunea repugnanţei” (cunoscută şi ca „factorul yuck”, Leon Kass, 1997), sau apeluri vagi la „bunul-simţ”. Deoarece avem de-a face cu minţi complet străine pentru noi, am putea spune extraterestre din punctul nostru de vedere, trebuie să fim expliciţi – da, chiar algoritmici.

Maşinile autonome, dronele, tranzacţiile la bursă, căutările Agenţiei Naţionale de Securitate etc. necesită decizii rapide, aprobate în prealabil. Putem face observaţii interesante în multe aspecte ale eticii pe care am încercat să le izolăm şi să le explicăm secole de-a rândul. Provocările au inclus priorităţi conflictuale, precum şi predilecţii cognitive întipărite la nivel biologic, sociologic şi semilogic. În mod notabil departe de consens în dogme universale despre drepturile omului se află noţiunile de confidenţialitate şi demnitate, chiar dacă acestea influenţează multe legi şi regulamente.

Oamenii ar putea avea dreptul de a citi (şi schimba) minţile computerelor pentru a vedea de ce iau decizii care intră în conflict cu instinctele noastre de Homo Sapiens. Nu este just ca maşinile să ceară acelaşi lucru de la noi? Observăm creşterea mişcărilor către transparenţă în conflictele financiare potenţiale; *softwa*re-urile „cu sursa deschisă”, *hardware* şi *wetwarey* Legea privind accesul echitabil la ştiinţă şi tehnologie; şi Open Humans Foundation.

În cartea sa din 1976, *Computer Power and Human Reason*, Joseph Weizenbaum argumenta că maşinile nu ar trebui să înlocuiască oamenii în situaţiile care necesită respect, demnitate sau îngrijire, în timp ce alţii (autoarea Pamela McCorduck şi informaticieni precum John McCarthy şi Bill Hibbard) au răspuns că maşinile pot fi mai imparţiale, calme şi consecvente, mai puţin abuzive sau răuvoitoaredecât oamenii în astfel de posturi.

Egalitate

La ce se referea tânărul Thomas Jefferson, în vârstă de treizeci şi trei de ani, în 1776, când scria „Noi menţinem aceste adevăruri a fi evidente de la sine, acela că toţi oamenii sunt creaţi egali, acela că sunt dotaţi de către Creator cu anumite drepturi inalienabile şi acela că printre aceste drepturi se află şi dreptul la viaţă, dreptul la libertate şi dreptul de a-şi căuta fericirea”? Spectrul oamenilor din prezent este vast. În 1776, termenul „oameni” nu-i includea pe oamenii de culoare sau pe femei. Chiar şi astăzi, oamenii născuţi cu probleme congenitale cognitive sau comportamentale sunt destinaţi unui tratament inegal (fie el chiar şi plin de compasiune, în unele cazuri) – sindromul Down,

boala Tay-Sachs, sindromul X fragil, paralizie cerebrală şi aşa mai departe.

Şi, pe măsură ce schimbăm locul geografic şi ne maturizăm, drepturile noastre inegale se schimbă spectaculos. Embrioni, bebeluşi, copii, adolescenţi, adulţi, pacienţi, infractori, identităţi de gen şi preferinţe de gen, cei mai bogaţi şi cei mai săraci – toate aceste entităţi au drepturi şi realităţi socioeconomice diferite. O cale către noi tipuri de minţi ce obţin şi păstrează drepturi similare elitelor umane ar fi să păstreze o componentă din genul Homo, cum ar fi un scut uman sau un monarh/CE-O reprezentativ, ceva semna orbeşte documente tehnice enorme şi va lua decizii financiare, de sănătate, diplomatice, militare sau de securitate pripite. Probabil ne va fi foarte greu să oprim, să modificăm sau să ştergem (să omorâm) un computer şi memoriile sale – în special dacă s-a împrietenit cu oamenii şi a făcut pledoarii spectaculos de convingătoare pentru supravieţuire (aşa cum ar face toţi cercetătorii excelenţi luptând pentru vieţile lor).

Chiar şi Scott Adams, creatorul lui *Dilbert*, şi-a spus părerea pe această temă, părere susţinută de experimentele de la Universitatea Eindhoven din 2005, care observă cât de susceptibili sunt oamenii la experimentele Milgram în care robotul este victimă, făcute la Yale, începând cu anul 1961. Date fiind multele drepturi ale corporaţiilor, inclusiv dreptul de a deţine proprietăţi, este probabil ca şi alte maşini să obţină drepturi similare şi va fi greu să administrăm inechităţile drepturilor selective pe gradienţi multiaxiali de intelect şi sentimente de înlocuire.

Reguli radical divergente pentru oameni faţă de nonoameni şi hibrizi

Împărţirea observată mai sus pentru variaţia din cadrul speciei Homo Sapiens în ceea ce priveşte drepturile explodează într-o revoltă de inegalităţi imediat ce trecem la entităţi care se suprapun (sau se vor suprapune în curând) peste spectrul umanităţii. Pe Google Stree View, feţele oamenilor şi numerele de înmatriculare are maşinilor sunt înceţoşate. Dispozitivele video sunt excluse din multe locuri, cum ar fi curţile de judecată şi şedinţele comitetelor. Camerele publice şi camerele care se pot purta pe corp cu softuri de recunoaştere facială intră în zona tabuurilor. Oare ar trebui excluşi şi oamenii care au sindromul hipertimestic sau memorii fotografice din aceste contexte?

Nu ar trebui, oare, ca oamenii cu prozopagnozie (incapacitatea de a identifica o faţă cunoscută) sau uitucii să beneficieze de softuri de recunoaştere facială şi recunoaştere optică a caracterelor oriunde merg şi, dacă ei au acest drept, de ce nu l-ar avea oricine? Dacă avem cu toţii aceste instrumente într-o anumită măsură, nu ar trebui să beneficiem cu toţii de ele?

Aceste scenarii sunt un ecou al nuvelei din 1961 a lui Kurt Vonnegut „Harrison Bergeron”, în care aptitudinile excepţionale sunt suprimate din respect pentru cel mai mic numitor comun al mediocrităţii societăţii. Experimente mentale cum ar fi Camera chinezească a lui John Searle şi Cele trei legi ale roboticii lui Isaac Asimov fac apel la tipurile de intuiţii care dau bătăi de cap creierelor omeneşti, aşa cum au demonstrat Daniel Kahneman, Amos Tversky şi alţii. Experimentul Camera chinezească postulează că o minte compusă din componente mecanice şi Homo Sapiens nu poate fi conştientă, indiferent cât de competentă ar fi la conversaţia (chinezească) umană inteligentă, dacă un om nu poate identifica sursa conştiinţei şi nu o poate „simţi”. Preferinţa forţată pentru prima şi a doua lege a lui Asimov favorizează minţile omeneşti în detrimentul oricărei alte minţi prezente ca o umbră ştearsă în a treia lege, cea a autoconservării.

Dacă roboţii nu au exact aceeaşi conştiinţă ca oamenii, atunci acest motiv este folosit ca scuză pentru a le conferi drepturi diferite, analog argumentelor conform cărora alte triburi sau rase sunt mai puţin umane. Roboţii chiar manifestă liberul-arbitru? Sunt ei deja conştienţi? Roboţii Qbo au trecut „testul oglinzii” de autorecunoaştere şi roboţii NAO un test corespunzător în care şi-au recunoscut propria voce, din care au dedus starea lor internă.

Pentru liberul-arbitru, avem algoritmi care nu sunt nici complet determinişti, nici aleatorii, ci au ca scop un proces probabilistic de luare a deciziilor aproape optim. Se poate argumenta că aceasta este o consecinţă darwmistă practică a teoriei jocurilor. În ce priveşte multe (nu toate) jocuri/probleme, dacă sunt complet previzibile sau complet aleatorii, atunci avem tendinţa să pierdem.

Dar care este atracţia liberului-arbitru? Din punct de vedere istoric, ne-a oferit o cale de a da vina în contextul recompensei şi pedepsei pe Pământ sau în viaţa de apoi. Obiectivele pedepsei pot include ajustarea priorităţii indivizilor de a ajuta la supravieţuirea speciei. În cazuri extreme, asta poate însemna privarea de libertate sau alte restricţii, dacă întărirea pozitivă / negativă skinneriană nu poate proteja societatea. În mod clar, astfel de instrumente se pot aplica liberului-arbitru, în sens larg – oricărei maşini al cărei comportament dorim să-l controlăm.

Putem să argumentăm dacă robotul experimentează cu adevărat *qualia* subiectivă a liberului-arbitru sau conştiinţei de sine, dar acelaşi lucru se aplică şi la evaluarea unui om. Cum ştim dacă un sociopat, un pacient în comă, o persoană cu sindromul Williams sau un bebeluş au acelaşi liber-arbitru sau aceeaşi conştiinţă de sine ca noi? Şi de ce ar conta, practic? Dacă oamenii (de orice fel) pretind în mod convingător să experimenteze conştiinţa, durerea, credinţa, fericirea, ambiţia şi/

sau utilitatea pentru societate, ar trebui, oare, să le negăm drepturi pentru că acele *qualia* ipotetice ale lor sunt ipotetic diferite de ale noastre?

Liniile drepte şi roşii ale prohibiţiei, peste care, teoretic, nu vom trece niciodată, par din ce în ce mai mult temporare şi neclare. Linia dintre oameni şi maşini nu mai este la fel de clară, atât pentru că maşinile devin mai omeneşti, cât şi pentru că oamenii devin mai maşinali – nu doar pentru că urmăm aproape orbeşte script-urile GPS, tweet-urile făcute din reflex şi marketingul ingenios conceput, ci şi pentru că digerăm din ce în ce mai multe informaţii cu creierul şi mecanismele noastre de programare genetică. Iniţiativa NIH BRAIN dezvoltă tehnologii inovatoare şi le foloseşte pentru a cartografia conexiunile şi activitatea circuitelor mentale, astfel încât să îmbunătăţească bunurile neurobiologice electronice şi sintetice.

Diverse linii roşii depind de excepţionalismul genetic, în care genetica este considerată o moştenire permanentă (deşi se poate dovedi că este reversibilă), în timp ce tehnologiile scutite (şi letale) de la aceasta, cum ar fi automobilele, sunt practic ireversibile din cauza forţelor sociale şi economice. În cadrul geneticii, o linie roşie ne face să interzicem sau să evităm alimentele modificate genetic, dar să îmbrăţişăm bacteriile modificate genetic care creează insulină sau oamenii modificaţi genetic – martori stau tratamentele aprobate în Europa pentru adulţi şi embrioni umani.

Linia pentru manipularea germenilor pare mai puţin sensibilă decât media, o linie practică trasă între siguranţă şi eficienţă. Căsătoriile între doi transportatori sănătoşi ai aceleiaşi boli genetice au de ales între a nu face copii, 25% şanse de a-şi pierde embrionul prin avort (spontan sau indus), 80% şanse de pierdere prin fertilizare in vitro sau, posibil, o pierdere potenţială de 0% a embrionului prin inginerie genetică a spermei (linia germinală). Pare prematur să declarăm această ultimă opţiune improbabilă.

Pentru cercetările „pe oameni", consultăm Declaraţia de la Helsinki, din 1964, aducându-ne aminte şi de experimentul cu sifilis Tuskegee 1932 – 72, posibil cea mai notorie cercetare biomedicală din istoria Statelor Unite. În 2015, Nonhuman Rights Project a intentat un proces la Curtea Supremă din New York în numele a doi cimpanzei păstraţi pentru cercetare de Universitatea Stony Brook. Decizia curţii de apel a fost că cimpanzeii nu trebuie trataţi ca nişte persoane din punct de vedere legal pentru că „nu au îndatoriri şi responsabilităţi în societate", în ciuda a ceea ce susţineau Jane Goodall şi alţii că, de fapt, ar avea, precum şi în ciuda argumentelor că o astfel de decizie ar putea fi aplicată şi copiilor sau persoanelor cu handicap.[69]

Ce împiedică extrapolarea acestei decizii şi asupra altor animale, organoide, maşini şi hibrizi? Promovând interdicţii asupra „armelor

autonome" (Hawking, Musk, Tallinn, Wilczek, Tegmark şi alţii), am demonizat un tip de maşină „neinteligentă", în timp ce alte maşini – de exemplu, cele compuse din mulţi Homo Sapiens care votează – pot fi mai letale şi mai prost îndrumate.

Transoamenii au populat deja Pământul? Să ne gândim la populaţiile care nu au avut contact cu alţi oameni, cum ar fi Sentinelese şi Andamanese din India, Korowai din Indonezia, Mashco-Piro din Perú, Pintupi din Australia, Surma din Etiopia, Ruc din Vietnam, Ayoreo-Totobiegosode din Paraguay, Himba din Namibia şi zeci de triburi din Papua-Noua Guinee. Cum ar răspunde acestea sau strămoşii noştri? Am putea defini ca „transumani" oamenii şi culturile pe care oamenii care trăiesc într-o cultură modernă, însă care nu este tehnologizată, nu i-ar putea înţelege.

Astfel de oameni moderni din Epoca de Piatră ar înţelege foarte greu de ce sărbătorim dovezile recente ale undei gravitaţionale LIGO care susţin teoria generală a relativităţii veche de o sută de ani. S-ar scărpina în cap gândindu-se de ce avem ceasuri atomice sau sateliţi GPS pentru a putea ajunge acasă sau cum şi de ce ne-am extins vederea de la o bandă optică restrânsă la spectrul complet de la radio la gamma. Ne putem mişca mai repede decât orice altă specie în viaţă; putem atinge viteza necesară pentru a scăpa de gravitaţia Pământului şi putem supravieţui în vidul rece al spaţiului.

Dacă aceste caracteristici (şi alte câteva sute) nu constituie transumanism, atunci ce? Dacă simţim că judecătorul transumanismului nu ar trebui să fie oameni paleoculturali, ci oamenii recenţi, atunci cum putem ajunge la transumanism? Noi, „oamenii recenţi", vom putea întotdeauna să înţelegem fiecare increment tehnologic nou – nefiind niciodată surprinşi îndeajuns pentru a declara atingerea unei ţinte transumane (în mişcare). Profetul science-fiction William Gibson spunea: „Viitorul este deja aici – doar că nu este distribuit uniform". Deşi acest lucru subestimează următoarea rundă a „viitorului", cu siguranţă milioane de oameni sunt deja transumani – majoritatea dintre noi dorind şi mai mult. Întrebarea „Ce a fost un om?" s-a metamorfozat deja în „Care au fost multele tipuri de transoameni?… Şi care au fost drepturile lor?"

Capitolul 24
Utilizarea artistică a fiinţelor cibernetice

Caroline A. Jones

Caroline A. Jones este profesor de istoria artei la Departamentul de Arhitectură de la MIT şi autoarea cărţilor Eyesight Alone: Clement Greenbergs Modernism Bureaucratization of the Senses, Machine in

the Studio: Constructing the Postwar American Artist şi The Global Work of Art.

Interesul lui Caroline A. Jones în arta modernă şi contemporană este îmbogăţit de disponibilitatea de a se cufunda în tehnologiile implicate în producerea, distribuirea şi recepţia ei. „În calitate de istoric al artei, multe dintre întrebările mele se referă la ce fel de artă putem face, ce tip de gânduri putem produce şi cel fel de idei putem găsi care-l pot împinge pe om dincolo de parametrii noştri încăpăţânaţi, egoişti, «concentraţi doar pe grupul nostru mic». Filosofii şi filosofiile de care sunt atrasă sunt cei care pun sub semnul întrebării obsesia occidentală a individualismului. Aceştia vin din foarte multe locuri diferite şi reînvie atât de multe întrebări şi probleme care au fost puse în anii 1960".

Recent, şi-a concentrat atenţia pe istoria ciberneticii. Cursul său de la MIT – Maşini automate, automatism, sisteme, cibernetică – explorează istoria interfeţei dintre om şi maşină în termeni de feedback, explorând latura culturală, şi nu pe cea inginerească, a acestei idei. Începe cu lecturi obligatorii din Wiener, Shannon şi Turing şi apoi pivotează de la oamenii de ştiinţă şi ingineri la lucrările şi ideile artiştilor, feministelor şi teoreticienilor postmoderni. Scopul ei: să ajungă la o paradigmă centrală nouă a evoluţiei care este bazată pe cultură – „comunalismul şi simbioza dintre specii, mai degrabă decât supravieţuirea celui mai bine adaptat individ".

În calitate de istoric, Caroline face diferenţa între ceea ce ea a denumit „cibernetica de stânga" şi „cibernetica de dreapta": „Ce înseamnă cibernetica de stânga? La un anumit nivel, este un joc de cuvinte: cibernetica «lăsată» în urmă[70]. La alt nivel, este o grupare politică vagă ce denotă coasta stângă: California, Esalen, grupul pe care Dave Kaiser îl numeşte «fizicienii hipioţi». Nu este un termen corespunzător, dar este o cale de a recunoaşte că a existat un grup controlat de complexul industrial militar, uneori în mod nefericit, care ne-a oferit instrumentele critice".

Arta cibernetizată este foarte importantă, dar arta pentru viaţă cibernetizată este mai importantă.

NAMJUNE PAIK 1966

Inteligenţa artificială nu este ce şi-au dorit iniţial artiştii de la cibernetică, atunci când a apărut cartea lui Norbert Wiener, *The Human Use of Human Beings*: *Cybemetics and Society*, în anul 1950. Gama de artişti care se identificau cu cibernetica în anii 1950 şi 1960 a avut iniţial puţin acces la „maşini de gândit". În plus, ingineri iscusiţi făcuseră deja roboţi ţestoase, jongleri şi roboţei care trăgeau la

lumină, dar nu şi creiere gigantice. Folosind conexiuni, fire de cupru, comutatoare simple şi senzori electronici, artiştii au mers pe urmele ciberneticienilor, creând sculpturi şi medii care simulau o conştiinţă interactivă – mişcări analoge şi interfeţe care aveau mai mult de-a face cu instinctele şi politicile sexuale de după război decât cu automatizarea producţiei de cunoştinţe. Obscurizată acum de o ideologie a „inteligenţei" libere, nepriponite de componente *hardware* sau de trup, inteligenţa artificială a uitat de primele zile ale asimilării ciberneticii de către artişti. Acele eforturi merită revizitate; au modelat relaţiile cu ceea ce filosofii francezi Gilles Deleuze şi Félix Guattari au denumit „*phylum*-ul maşinilor", care are de-a face cu modul în care oamenii gândesc şi simt în corpurile conectate la o lume fizică, materială, stimulantă emoţional şi semnalizatoare.

Cibernetica se pare că s-a prăbuşit acum într-un discurs atotpătrunzător despre inteligenţa artificială care era departe de a fi predestinată. Cuvântul „cibernetică" a căpătat o prospeţime nouă pentru concepte care aveau cel puţin patru secole: noţiuni despre *feedback*, amortizare automată, homeostază biologică, calcul logic şi gândire sistemică ce au existat încă din Iluminism (fiind sporite de Revoluţia industrială). Numele din această descendenţă îi includ pe Descartes, Leibniz, Sadi Carnot, Clausius, Maxwell şi Watt. Însă numele dat de Wiener a avut efecte culturale profunde. Ubicuitatea în zilele noastre a prefixului ciberconfirmă dorinţa de semnificant nou pentru relaţiile întortocheate dintre oameni şi maşini. Cu sensul pe care l-a folosit Wiener, lucrurile „ciber" implicau pur şi simplu „control şi comunicare între animal şi maşină". Dar, după revoluţia digitală, „ciber" a trecut dincolo de servomécanisme, cicluri de *feedback* şi comutatoare, înglobând noţiuni de *software*, algoritmi şi cyborgi. Activitatea artiştilor înclinaţi către cibernetică priveşte comportamentele emergente ale vieţii care scapă inteligenţei artificiale în condiţia sa curentă.

În ceea ce priveşte sensul iniţial, Wiener a ajuns până la greaca antică pentru a împrumuta cuvântul pentru „timonier" (*kubernétés*), o figură masculină care canalizează putere şi instinct la timona navei, care citea valurile, judeca vântul, ţinea o mână pe cârmă şi îi direcţiona pe sclavi când trăgeau fără discernământ (mecanic) la vâsle. Cuvintele greceşti migraseră deja în engleza modernă prin latină, trecând de la kuberla *guber-y* rădăcina cuvintelor „guvernamental" şi „guvernator", alt termen pentru controlul masculin, dezvoltat de James Watt pentru a-şi descrie dispozitivul său din secolul al XIX-lea pentru modularea unui motor cu aburi impetuos. Cibernetica a preluat astfel idei care făceau de mult timp analogii între oameni şi dispozitive şi le-a generalizat într-o ştiinţă aplicată, adăugând particula „— ică". Cei trei „*e*" ai lui Wiener (comandă,

control, comunicare) se bazau pe matematica probabilităţii pentru a formaliza sisteme (biologice sau mecanice) teoretizate ca un set de intrări de informaţii care obţine ieşiri sub formă de acţiune într-un mediu – o agendă musculoasă şi trupeşă, adesea minimalizată în genealogiile inteligenţei artificiale.

Dar etimologia nu reuşeşte să exprime suficient entuziasmul resimţit de participanţi, când matematica s-a unit cu biologia teoretică (Arturo Rosenblueth) şi teoria informaţiei (Claude Shannon, Walter Pitts, Warren McCulloch), pentru a produce un baraj de cercetări şi publicaţii interdisciplinare văzute ca prezentând posibilitatea de a schimba nu doar modul în care se face ştiinţă, ci şi modul în care oamenii din viitor vor interacţiona cu tehnosfera. În cuvintele lui Wiener, „ne-am modificat mediul atât de radical încât trebuie acum să ne modificăm pe noi înşine pentru a putea exista”[71]. Întrebarea presantă este: cum ne modificăm? Mergem în direcţia potrivită sau ne-am pierdut calea, devenind instrumentele instrumentelor noastre? Revizitarea istoriei recente a contribuţiei umaniştilor/artiştilor la cibernetică ne-ar putea ajuta să ne orientăm către un viitor mai puţin periculos şi mai etic.

Anul 1968 a fost un vârf al difuziei culturale şi al asimilării artistice a termenului. În acel an, galeria Howard Wise şi-a deschis expoziţia de „sculptură cibernetică” a lui Wen-Ying Tsai în centrul Manhattanului, iar emigranta politică poloneză Jasia Reichardt şi-a deschis expoziţia „Serendipitatea cibernetică” la Institutul de Artă Contemporană din Londra (cuvântul „cibernetic” din titlul ei voia să însemne „făcut cu sau de un computer”, chiar dacă majoritatea operelor de artă expuse nu aveau computere concrete în circuitele lor). Cele două decenii dintre 1948 şi 1968 au reprezentat atât răspândirea conceptelor cibernetice într-o cultură mai largă, cât şi răspândirea maşinilor de calcul într-o migrare lentă, de la echipamente militare brevetate până la corporaţiile multinaţionale şi laboratoarele universitare, unde artiştii au început să primească acces. Disponibilitatea componentelor cibernetice – „organe senzoriale” (ochi electronici, senzori de mişcare, microfoane) şi „organe făptuitoare” (conexiuni electronice, comutatoare, sisteme hidraulice şi pneumatice) – pe frontul hobby-urilor de acasă făcea din computer mai puţin un „creier electronic” decât un organ ajutător dintr-un set de componente. Nu exista încă o metaforă predominantă pentru „inteligenţa artificială”. Aşadar artiştii erau deja bricoleuri de corpuri electronice, interesaţi mai mult de acţiuni decât de calcule sau cogniţie. Existau indicii care făceau din „computere” calculatoare în cursa către *Homo rationalis*, dar mai mult în aspiraţie decât în realizare.

În lumina convergenţei digitale de astăzi a instrumentelor de

imagistică artistică/ştiinţifică, spectacolul lui Reichardt era profetic în insistenţa sa în ceea ce priveşte confundarea graniţelor dintre artă şi ceea ce ar putea fi numită „ştiinţă creativă aplicată". Conform catalogului, „niciun vizitator al expoziţiei, dacă nu citeşte toate notele referitoare la toate lucrările, nu va şti dacă priveşte ceva făcut de un artist, inginer, matematician sau arhitect". Astfel, robotul hilar de disfuncţional al lui N-am June Paik, *Robot K-456* (1964), prezentat pe coperta catalogului şi descris ca „un robot feminin cunoscut pentru comportamentul său deranjant şi idiosincratic", se va întrece cu un *Colloquy of Mobiles* (1968), coregrafic construit de ciberneticianul mai puţin cunoscut Gordon Pask. Pask a lucrat cu un scenograf de teatru londonez pentru a construi un aparat „masculin" fusiform cu balamale şi tije, configurat să comunice cu entităţile „feminine" bulboase din fibră de sticlă care se aflau prin preajmă. Dacă cineva ar putea să cartografieze caracteristicile programului (sau să spicuiască teatrul său reacţionar de gen) fără a citi eseul din catalog este o întrebare deschisă. Semnificativă este concentrarea lui Pask pe comportamentele automatelor sale, pe interactivitatea lor, pe responsivitatea lor într-un mediu modulat artificial şi pe „reflectarea" comportamentului uman.

„Serendipitatea cibernetică" de la Institutul de Arte Contemporane a introdus o paradigmă importantă: ecosistemul maşinic, în care privitorul este o parte biologică, având ca sarcină să-şi dea seama care sunt declanşatorii interacţiunii. Vizitatorii din acele galerii londoneze au devenit brusc „organisme cibernetice" – cyborgi –, deoarece, pentru a experimenta corespunzător arta, trebuia să intri într-o conversaţie simbiotică cu servomecanismele. Această orientare către mediile interactive dintre om şi maşină ca estetică devine din ce în ce mai clară când examinăm câteva alte opere de artă ale perioadei, începând cu una care constituie un exemplu timpuriu al comportamentului emergent – *Senster*, sculptura interactivă făcută de artistul/inginerul Edward Ihnatowicz (1970), celebrată de inginerul în robotică medicală Alex Zivanovic, editor al site-ului web dedicat carierei puţin cunoscute a lui Ihnatowicz, ca „una dintre primele opere de artă robotice interactive controlate de computer". Aici „computerul" îşi face intrările sale (deşi este un dispozitiv limitat, cu doisprezece biţi). Dar mai degrabă decât „inteligenţa", Ihnatowicz a urmărit un avatar al comportamentului afectiv. Cheia succesului neaşteptat al lui *Senster* a fost programarea cu care Ihnatowicz făcea aparatul hidraulic de cinci metri (designul său cu balamale şi aspect de miraj era inspirat de un cleşte de homar) să transmită sfială când răspundea oamenilor din proximitate. Canalele de sunet şi senzorii de mişcare ai lui *Senster* erau setaţi să se tragă în spate la zgomote puternice sau mişcări bruşte, aparent agresive. Doar oamenii care erau

dispuşi să vorbească blând şi să-şi moduleze gesturile urmau să fie recompensaţi de abordarea liniştită şi curioasă a lui *Senster* – o experienţă care a devenit reală şi pentru Ihnatowicz însuşi, când acesta a structurat iniţial programul, iar maşina s-a întors către el cu solicitudine, după ce a tuşit pentru a-şi elibera gâtul.

În aceste utilizări artistice ale fiinţelor cibernetice, simţim o necesitate crescândă de a instrui publicul în a se resimţi pe sine ca integrat într-un mediu tehnologizat, modificându-se pe sine pentru a comunica intuitiv cu maşinile. Această necesitate a devenit deja explicită în expoziţia de „Sculptură Cibernetică" a lui Tsai. Cei care experimentează instalaţia lui în care se cufundă cu totul se aşteptau să experimenteze cu viaţa maşinică: ce comportament ar declanşa servomecanismele? Asemenea, participantul din galeria umană ar fi trebuit să explice protocolul: „Bate din palme – asta face sculpturile să răspundă". În cuvintele unui critic de atunci:

> *Un schelet de tije din oţel se ridică de pe o placă. Această bază vibrează cu treizeci de cicluri pe secundă; tijele flexează rapid, în curbe armonice. Amplasat într-o cameră întunecată, este iluminat de lumini stroboscopice. Pulsul luminilor intermitente variază – sunt conectate la senzori de sunet şi proximitate. Rezultatul este că, atunci când cineva se apropie de un Tsai sau face un zgomot în vecinătatea lui, obiectul răspunde. Tijele par să se mişte; se observă o sclipire, o lumină şi un balet straniu de metal, ale cărui mişcări variază de la nemişcat la tremurat şi apoi înapoi la o ondulaţie lentă, inefabil de senzuală.*

La fel ca *Senster* aparatul stimula (şi simula) o interacţiune afectivă, şi nu una raţională. Oamenii simţeau că întâlnesc comportamente ce indică viaţa sensibilă; entităţile lui Tsai era adesea clasificate ca „vegetale" sau „acvatice". Astfel de ambiţii ce ţin de mediu şi de mişcare erau notorii în lumea internaţională a artei din vremea aceea. Dincolo de grajdul de la Howard Wise, erau emigranţii politic care formau colectivul GRAV în Paris, „arhitecturile cibernetice" ale lui Nicolas Schoffer, giraţiile de lumină şi plastic ale German Zero Gruppe şi altele – toate definind şi informând genul de artă cibernetică ce se anunţa.

Utilizarea artistică a fiinţelor cibernetice de la sfârşitul anilor şaizeci nu a făcut nicio investiţie în „inteligenţă". Ştiind că maşinile sunt lipsite de inteligenţă şi incapabile de emoţie, aceşti creatori erau încrezători când puneau în scenă simulări oneste. Ce-i interesa pe ei erau acţiuni, instincte şi afecte ce evocau mişcări ale maşinii; imitau comportamente sexuale şi animalice, ca şi cum ar fi sub pragul conştiinţei. Astfel de artişti nu erau interesaţi de manipularea datelor sau informaţiilor (deşi Hans

Haacke se va fi îndreptat în acea direcţie în 1972 cu lucrările sale de „sisteme în timp real”). Cultura cibernetică pe care artiştii şi cercetătorii o configurau pe două continente integra omul în tehnosferă şi seducea percepţia cu comportamentele graţioase şi sensibile ale filonului maşinic. „Artificialul” şi „naturalul” erau întreţesute în această estetică cibernetică timpurie.

Dar nu se va termina aici. Crucială pentru extinderea acestui set necritic şi în mare parte masculin de medii cibernetice se va dovedi a fi o cohortă critică radicală de artiste uimitoare care vor apărea în anii 1990, pe deplin conştiente de predecesorii lor în artă şi tehnologie, dar poate mai mult inspirate de fondatoarele feministe ale revistei din 1970 *Radical Software* şi boomul cultural al inspiratei polemici din 1984 a Donnei Haraway, „Manifestul cyborgului”. Teatrul scârţâitor de gen al lui Paik şi Pask, creaturile inocente ale lui Ihnatowicz şi Tsai au fost mobilizate ca isteţe, performative şi postmoderne, ca în lecţia lui Lynn Hershman „Dollie Clone Series” (1995 – 1998), care consta în asamblajele interactive *Cybe Roberta* şi *Tilliey păpuşa telerobotică*, lucrând în tehnosferă cu profesionalismul burlescului, făcându-ne cu ochiul şi pliindu-ne pe noi, privitorii, într-o conştiinţă explicită a poziţiei noastre voyeuriste, atât în calitate de subiecţi care privesc, cât şi ca obiecte de privit.

Tehnosferă „inocentă” stabilită de sculptorii cibernetici bărbaţi din anii 1960 a fost, până în anii 1990, identificată de artistele feministe drept o condiţie complet înecătoare ce ne solicită atenţia critică. În acelaşi timp, feministele au abordat problema cărei „inteligenţe” încearcă să o simuleze inteligenţa artificială. Pentru o artistă ca Hershman Leeson, răspunzând la „triumful” tehnic de a o clona pe oaia Dolly, era crucial să facem conexiunea între producţia de carne şi „maşinile de came”. Hershman Leeson producea „păpuşi” drept clone, oferind o încadrare critică a modului în care individualizarea contemporană a devenit o parte din domeniul logic, replicativ şi plastic.

Deşi tehnofeministele din anii 1990 şi 2000 nu erau „ciber” tot timpul, lucrările lor au complicat calităţile dominante maşinice şi cinetice ale mediilor tehnologice anterioare ale artiştilor bărbaţi. Telecyborgul androgin din *Imaginaţiony Dead Imagine* (1991) al lui Judith Barry, de exemplu, nu avea nicio parte în mişcare. El/ea era compus din semnale pure, proiecţii intermitente pe suprafeţe plate. În lucrarea sa, Barry a comentat asupra efectelor alienante ale tehnologiei secolului XX. Imaginea unui cap androgin umple un cub enorm făcut din ecrane de trei metri pe cinci părţi, montate pe o bază de oglinzi tot de trei metri. O diversitate de lichide vâscoase şi care nu arată prea bine (galbene, roşietice-portocalii, maro), materiale uscate (rumeguş? fainã?) şi chiar insecte îşi croiesc drum pe cap, a cărui

sublimitate stoică este minunat virtualizată pe ecranele enorme ale lucrării. *Dead Imagine*, prin forma sa „platonică” la scară mare şi cubică, rămâne atât artificială, cât şi blocată în corp – refuzând o „inteligenţă” detaşată ca nefiind deloc inteligenţă.

Artiştii din noul mileniu moştenesc această tradiţie critică şi locuiesc în paradigmele curente ale inteligenţei artificiale, care au alunecat de la simulări parţiale la pretenţii de inteligenţă. În propunerea din 1955, care este considerată a fi prima utilizare tipărită a sintagmei „inteligenţă artificială”, informaticianul John McCarthy şi colegii săi Marvin Minsky, Nathaniel Rochester şi Claude Shannon au emis ipoteza că „toate aspectele învăţării sau orice altă caracteristică a inteligenţei pot fi, în principiu, atât de precis descrise încât o maşină le poate simula”. Acest obiectiv teoretic modest s-a umflat în ultimii şaizeci şi patru de ani, fiind exprimat acum de Google Deep Mind ca o ambiţie de a „rezolva inteligenţa”. De a sparge codul! Dar, din păcate, ce auzim crăpând nu este codul, ci capitalismul la scară mică, contractul social şi structura politeţii. Depăşirea meseriilor de şofer de taxi şi de camion, robotizarea marketingului direct, hegemonizarea divertismentului, privatizarea utilităţilor şi depersonalizarea îngrijirii medicale – sunt acestea „bicele” de care Wiener se temea că vom învăţa să le iubim?

Artiştii nu pot rezolva niciuna dintre aceste probleme. Dar ne pot reaminti de potenţialul creativ al căilor nebătătorite – intersecţiile care ne ieşeau în cale în anii 1970, înainte ca „informaţia” să devină capital şi „inteligenţa” să fie echivalată cu colectarea de date. Puternic evocator a ce se poate face cu instrumentele contemporane când revizităm posibilităţile din trecut este „licuriciul” artistului francez Philippe Parreno, numit astfel pentru a evita repetiţia numelui său adevărat: *Cu un comportament instinctiv ritmic pentru a putea să călătoreşti dincolo de forţele existente ale vieţii* (2014). Descrisă de artist ca un „automat”, instalaţia sculpturală juxtapune o proiecţie intermitentă cu desene alb-negru de licurici cu o bandă de figuri binare oscilante, având culoarea verde pe negru. Desenele pe figuri binare sunt animate folosind algoritmii matematicianului John Horton Conway, folosiţi în Jocul Vieţii din 1970, „automatul celular”.

Conway a configurat parametrii astfel încât orice pătrat („celulă”) să fie aprins („viu”) sau stins („mort”) într-o matrice bidimensională infinită. Regulile sunt rezumate după cum urmează: o celulă individuală va muri rapid de singurătate. Dar o celulă care atinge trei sau mai multe celule „vii” va muri şi ea „din cauza aglomeraţiei”. O celulă supravieţuieşte şi prosperă dacă are doar doi vecini… şi aşa mai departe. Când o celulă moare, poate crea condiţiile necesare supravieţuirii altor celule, creând tipare care par să se mişte şi să crească, aprinzându-se în matrice ca impulsurile neuronale

evanescente sau grupuri bioluminescente de diatomee. În filmul din 2012 al lui Stephen Hawking *The Meaning of Life*, naratorul descrie modelul matematic al lui Conway ca simulând „cum ceva complex aşa cum este mintea ar putea lua naştere dintr-un set elementar de reguli", revelând ambiţiile înfumurate ce caracterizează inteligenţa artificială contemporană: „Aceste proprietăţi complexe iau naştere din legi simple ce nu conţin concepte cum ar fi mişcarea sau reproducerea", însă produc „specii", şi celulele „se pot chiar reproduce, la fel cum face viaţa în lumea reală".[72]

La fel cum face viaţa? Artiştii cunosc linguşirile simulărilor şi reprezentării, diferenţa dintre geniul artificiului şi realităţile a „ceea ce face viaţa". Lucrarea lui Parreno este un ansamblu intuitiv pentru modul în care experimentăm noi „viaţa" prin angajamente încorporate, perspectiviste. Conştiinţa noastră este întreţesută electric (cibernetic), însă nu răspundem ca şi cum acest set de simulări elegante generat de oameni ar avea propria sa *inteligenţă*.

Utilizarea artistică a fiinţelor cibernetice ne reaminteşte, de asemenea, că şi conştiinţa însăşi nu este doar „aici". Intră şi iese, armonizând aceste semnale senzoriale sclipitoare. Mintea se întâmplă mult în afara limitelor craniului (şi simulacrului acestuia, „placa de bază"). Cum spunea Mary Catherine Bateson, parafrazând conceptul tatălui său, Gregory, de cibernetică de rangul doi, mintea este un material „nu neapărat definit de o limită precum o bășică de piele"[73]. Parreno împerechează simulările artei cu simulările matematicii pentru a forţa ideea de sorginte wieneriană că un astfel de model nu este, în sine, la fel ca viaţa. Modele sunt doar părţi ale sistemelor de semnalizare ce constituie „inteligenţă" doar atunci când omoloagele lor biologice interacţionează cu ele în creare de sens. Inteligenţa artificială contemporană s-a marginalizat instrumentalizând şi particularizând sarcini şi subrutine, confundând aceste exerciţii cu înţelepciunea veritabilă. Scurta istorie culturală oferită aici ne reaminteşte că a privi datele ca pe o formă de inteligenţă, reţelele digitale ca „neuronale" sau indivizii izolaţi ca unităţi de viaţă este o metodă străină chiar şi pentru simulările brute ale lui Conway.

Putem stigmatiza aroganţa căpoasă a inteligenţei artificiale curente precum „cibernetica de dreapta", calea care a condus la sistemele de armament automatizate pe care le avem în prezent, ostilitatea prost deghizată a companiei Über pentru şoferii umani şi visurile capitaliste ale companiei Google. Acum trebuie să ne întoarcem la cibernetica de stânga – teoreticieni biologi şi antropologi încercând să înţeleagă sistemele inteligente independent de specie. Observaţia lui Gregory Bateson conform căreia corporaţiile nu fac decât să simuleze „agregate de părţi de persoane", cu decizii de maximizare a profitului tăiate din „părţi ale minţii mai mari şi mai

înţelepte", nu a fost niciodată mai oportună.[74]

Epistemologia cibernetică oferită aici sugerează o abordare nouă. Mintea individuală este imanentă, nu doar pentru corp, dar şi pentru căi din afara corpului, şi există o Minte mai mare, pentru care mintea individuală este doar un subsistem. Această Minte mai mare, susţine Bateson, este comparabilă cu Dumnezeu şi este, probabil, ceea ce unii oameni înţeleg prin conceptul de Dumnezeu, dar tot este imanentă sistemului social total interconectat şi ecologiei planetare. Aceasta nu este iluzia colectivă a unui „Dumnezeu" exterior, care vorbeşte din afara conştiinţei umane (această idee cu tradiţie lungă de monoteism, sugerează Bateson, conduce la viziuni despre natură şi mediu ca fiind, de asemenea, în afara omului „individual", făcând din ele „daruri bune de exploatat"). Mai degrabă, „Dumnezeul" lui Bateson este un locţiitor al experienţei noastre evanescente de interacţiune a conştiinţei în lume: Mintea mai mare, ca rezultat al intrărilor şi acţiunilor care devin apoi intrări pentru alte acţiuni împreună cu alte entităţi – pânze de relaţii simbiotice ce formează tipare cu care trebuie să armonizăm şi să simţim urgent.[75]

De la Tsai în 1970 până la Hershman Leeson în 1990 şi Parreno în 2014, artiştii au criticat cibernetica de dreapta şi s-au jucat cu experimente cibernetice alternative, întrupate, de mediu ale inteligenţei „artificiale". Utilizarea lor artistică a fiinţelor cibernetice oferă înţelepciunea simbioţilor experimentată în tipurile de *poiesis* care pot fi obţinute în această lume: ritmuri de semnale şi acţiuni intuitive ce produc mişcările vieţii în parteneriat cu tehnosfera electromecanică şi electromagnetică. Viaţa, în misterioasele sale îmbinări negentropice cu materia şi Mintea.

Capitolul 25
Inteligenţa artificială şi viitorul civilizaţiei

Stephen Wolfram

Stephen Wolfram este om de ştiinţă, inventator, fondatorul şi CEO-ul companiei Wolfram Research. Este creatorul programului de tehnică de calcul simbolică Mathematica şi al limbajului său de programare, limbajul Wolfram, precum şi al motorului de cunoştinţe Wolfram Alpha. De asemenea, este autorul cărţii A New Kind of Science. *În cele ce urmează, veţi găsi o transcriere editată dintr-un interviu live cu el realizat în decembrie 2015.*

În aproape patru decenii, Stephen Wolfram a fost un pionier în dezvoltarea şi aplicarea gândirii computaţionale, fiind responsabil

de multe inovaţii în ştiinţă, tehnologie şi afaceri.

Lucrarea sa din anul 1982 Maşinile automate celulare ca sisteme simple cu autoorganizare, scrisă la vârsta de douăzeci şi trei de ani, a fost prima dintre numeroasele sale contribuţii ştiinţifice semnificative care au avut ca scop înţelegerea originilor complexităţii în natură.

În această perioadă, Stephen a intrat o scurtă perioadă în viaţa mea. Tocmai înfiinţasem The Reality Club, o adunare neoficială de intelectuali, care s-au întâlnit în New

York pentru a-şi prezenta lucrările în faţa colegilor de breaslă din alte discipline. (Observaţie: în 1996, The Reality Club a ajuns online pe site-ul Edge.org.) Primul nostru vorbitor? Stephen Wolfram, un „wunderkind”[76] care a ajuns la Princeton, la Institutul de Studii Avansate. Îmi amintesc maniera sa concentrată când s-a aşezat pe o canapea din sufrageria mea şi a vorbit fără întrerupere aproximativ o oră în faţa grupului strâns.

Din acel moment, Stephen are ca scop să facă toate cunoştinţele lumii uşor computabile şi accesibile. Programul său, Mathematica, este sistemul definitoriu pentru tehnica de calcul modernă. Wolfram Alpha calculează răspunsuri la nivel de expert utilizând tehnologia inteligenţei artificiale. Consideră că limbajul său de programare Wolfram este primul limbaj adevărat de comunicare computaţională pentru oameni şi inteligenţe artificiale.

M-am reîntâlnit cu el cu patru ani în urmă, în Cambridge, Massachusetts, pentru o discuţie liberă pe tema inteligenţei artificiale. Stephen a intrat, m-a salutat, s-a aşezat şi, privind camera video aranjată ca să înregistreze conversaţia pentru Edge, a început să vorbească şi nu s-a oprit timp de două ore şi jumătate.

Eseul care urmează este o versiune editată a acelei sesiuni, care a fost un fel de curs aprofundat despre Wolfram, fiind un mod corespunzător de a încheia acest volum – la fel cum prelegerea lui Stephen de la Reality Club din anii 1980 a fost un mod minunat de a iniţia întreprinderea intelectuală în desfăşurare al cărei rezultat este comunitatea bogată de gânditori care-şi prezintă activitatea unul altuia şi publicului în această carte.

Eu văd tehnologia ca fiind automatizarea obiectivelor omeneşti pentru a le face executabile pentru maşini. Obiectivele omeneşti din trecut au presupus mutarea obiectelor dintr-un loc în altul cu ajutorul unui stivuitor, şi nu al mâinilor. Acum munca pe care o putem face automat, cu maşinile, este mai mult mentală decât fizică. Este evident că putem automatiza multe dintre sarcinile pe care noi, oamenii, ne mândrim de mult timp că le-am făcut singuri. Care este viitorul condiţiei umane în această situaţie?

Oamenii vorbesc despre viitorul maşinilor inteligente şi dacă vor prelua controlul şi vor hotărî ce să facă ele însele. Dar inventarea obiectivelor nu este ceva ce are o cale în automatizare. Cineva sau ceva trebuie să definească aceste obiective ale unei maşini – ce încearcă să execute. Cum sunt definite obiectivele? Pentru om, au tendinţa să fie definite de istoria personală, de mediul cultural, de istoria civilizaţiei noastre. Obiectivele sunt unic omeneşti. În ceea ce priveşte maşina, noi îi conferim un scop atunci când o construim.

Ce fel de lucruri posedă inteligenţă, obiective sau scopuri? Momentan, cunoaştem un exemplu mare şi lat, şi anume noi – creierele noastre, inteligenţa noastră omenească. Inteligenţa omenească, presupuneam eu cândva demult, este mult peste orice altceva există natural în lume; este rezultatul unui proces elaborat de evoluţie, evidenţiindu-se astfel de restul existenţei. Dar am realizat, ca urmare a ştiinţei pe care am practicat-o, că nu aşa stau lucrurile.

Oamenii pot spune, de exemplu, „Vremea are propria ei minte”. Aceasta este o afirmaţie animistă şi pare să nu aibă ce căuta în gândirea ştiinţifică modernă. Dar nu este atât de neserioasă pe cât sună. Ce face creierul omenesc? Un creier primeşte anumite intrări senzoriale, calculează, determină înfăptuirea anumitor acţiuni şi generează un anumit rezultat. La fel ca vremea. Tot soiul de sisteme, de fapt, calculează – indiferent dacă este vorba despre un creier sau, să spunem, un nor care răspunde la mediul său termic.

Putem să argumentăm că facem calcule mult mai sofisticate cu creierele noastre decât cele care apar în atmosferă. Dar se pare că există o echivalenţă mare între tipurile de calcule pe care le fac diversele tipuri de sisteme. Acest lucru pune într-o nouă lumină problema condiţiei umane, pentru că se pare că nu suntem atât de speciali pe cât credeam. Sunt toate aceste sisteme ale naturii care sunt, practic, echivalente în ceea ce priveşte capacităţile de calcul.

Ce ne face diferiţi de toate celelalte sisteme sunt elementele particulare ale istoriei noastre, care ne dau noţiunile pe care le avem noi de scop şi obiectiv. Acesta este un alt fel de a spune că, deşi cutia de pe biroul nostru gândeşte la fel de bine precum creierele noastre, nu va avea niciodată, intrinsec, obiective şi scopuri. Acestea sunt definite de trăsăturile noastre particulare – de biologia, psihologia şi istoria noastră culturală.

Când ne gândim la viitorul inteligenţei artificiale, trebuie să ne gândim la obiective. Cu ele contribuie oamenii; cu ele contribuie civilizaţia noastră. Execuţia acestor obiective este ceea ce putem automatiza din ce în ce mai bine. Care va fi viitorul oamenilor într-o astfel de lume? Ce rămâne de făcut atunci? Unul dintre proiectele mele a fost să înţeleg evoluţia scopurilor omeneşti în timp. Astăzi avem tot felul de scopuri. Dacă privim înapoi cu o mie de ani, oamenii

aveau cu totul alte obiective: cum fac rost de mâncare? Cum mă apăr? În lumea occidentală modernă, în cea mai mare parte, nu petrecem prea mult timp din viaţa noastră gândindu-ne la aceste scopuri. Din punctul de vedere al unui om de acum o mie de ani, unele obiective pe care le au oamenii din ziua de astăzi ar părea complet bizare – de exemplu, alergatul pe bandă. Acum o mie de ani, ar fi fost nebunie curată.

Ce vor face oamenii în viitor? Multe dintre scopurile pe care le avem astăzi sunt generate de un anumit tip de raritate. Resursele din lume sunt limitate. Oamenii vor să obţină mai multe dintr-o anumită categorie. Până şi timpul este puţin în viaţa noastră. În cele din urmă, acele forme de raritate vor dispărea. Cea mai pregnantă discontinuitate va fi cu siguranţă când vom obţine efectiv imortalitatea omenească. Dacă aceasta va fi obţinută biologic sau digital nu ştim încă, dar o vom obţine inevitabil. Multe dintre obiectivele noastre au la bază cel puţin parţial, mortalitatea noastră: „Voi trăi doar o perioadă de timp, aşa că ar fi bine să obţin asta şi asta”. Şi ce se întâmplă când majoritatea obiectivelor noastre sunt executate automat? Nu vom avea aceleaşi motivaţii pe care le avem astăzi. O întrebare la care mi-aş dori un răspuns este: ce vor alege să facă cu ele însele derivatele oamenilor din viitor? Unul dintre potenţialele rezultate negative este să joace jocuri video tot timpul.

Termenul „inteligenţă artificială” evoluează în utilizarea sa în limbajul tehnic. În prezent, inteligenţa artificială este foarte populară şi oamenii înţeleg oarecum ce înseamnă. Atunci când computerele abia fuseseră inventate, în anii 1940 şi 1950, titlul tipic dintr-o carte sau dintr-un articol de ziar despre computere era „creiere electronice uriaşe”. Ideea era că, la fel cum buldozerele şi motoarele cu aburi au automatizat munca mecanică, computerele vor automatiza munca intelectuală. Promisiunea s-a dovedit a fi mai greu de ţinut decât se aşteptau mulţi oameni. La început, exista mult optimism; mulţi bani din trezoreria de stat erau cheltuiţi pe astfel de eforturi la începutul anilor 1960. Pur şi simplu, nu funcţionau.

Există multe portretizări cu iz science-fiction ale computerelor din vremea aceea. Unul simpatic este *Desk Set*, despre un computer de tip IBM instalat într-o companie de difuzare a presei şi care îi lăsa pe toţi fără locuri de muncă. Este simpatic, deoarece computerului i se pun câteva întrebări de bibliotecă. Când eu şi colegii mei construiam Wolfram Alpha, una dintre ideile pe care le-am avut a fost să-l facem capabil să răspundă la toate acele întrebări de bibliotecă din *Desk Set*. În 2009, putea deja să răspundă la toate.

În 1943, Warren McCulloch şi Walter Pitts au introdus un model pentru cum ar putea funcţiona creierele conceptual şi formal – o reţea

neuronală artificială. Au văzut că modelul lor bazat pe creier ar face calcule la fel ca o maşină Turing. Din activitatea lor a reieşit că putem construi reţele neuronale asemănătoare creierului care ar acţiona ca nişte computere generale. Şi, de fapt, activitatea practică desfăşurată de angajaţii ENIAC, John von Newmann şi alţii pe calculatoare a venit pe linie directă nu din maşinile Turing, ci prin această cotitură a reţelelor neuronale.

Dar reţelele neuronale simple nu au realizat multe. Frank Rosenblatt a inventat un dispozitiv de învăţare pe care l-a numit „perception", care era o reţea neuronală cu un singur strat. La sfârşitul anilor 1960, Marvin Minsky şi Seymour Papert au scris o carte intitulată *Perceptions*, în care dovedeau practic că percepţionii nu ar putea face nimic interesant, ceea ce este corect. Percepţionii puteau să facă doar distincţii liniare între lucruri. Prin urmare, ideea a fost mai mult sau mai puţin abandonată. Oamenii spuneau: „Aceşti băieţi au scris că reţelele neuronale nu pot face nimic interesant, deci nicio reţea neuronală nu poate face nimic interesant, aşa că haideţi să lăsăm baltă reţelele neuronale". Această atitudine a rezistat ceva vreme.

Între timp, au mai apărut şi alte abordări ale inteligenţei artificiale. Una era bazată pe o înţelegere, la nivel formal, simbolic, a felului în care funcţionează lumea; alta era bazată pe statistici şi probabilităţi. Cu privire la inteligenţa artificială simbolică, unul dintre cazurile-test a fost: putem învăţa un computer să rezolve integrale? Putem învăţa un calculator să rezolve operaţii aritmetice? Existau sarcini de traducere automată, despre care oamenii credeau că ar fi un bun exemplu pentru ce pot face computerele. Ideea este că, până în anii 1970, acea abordare a picat.

Apoi a existat o tendinţă către dispozitive numite sisteme expert, care au luat naştere la sfârşitul anilor 1970 şi începutul anilor 1980. Ideea era ca o maşină să înveţe regulile pe care le foloseşte un expert, dându-şi seama astfel ce are de făcut. Şi această idee a pierit încet-încet. Apoi inteligenţa artificială a devenit ceva neserios de urmărit.

Am fost interesat de cum se poate construi o maşină asemănătoare inteligenţei artificiale de când am fost copil. Eram interesat în mod special de modul în care se pot lua cunoştinţele pe care le-am acumulat noi, oamenii, în civilizaţia noastră şi se poate automatiza răspunsul la întrebări în baza acelor cunoştinţe. M-am gândit cum s-ar putea realiza asta simbolic, construind un sistem care ar putea defalca întrebările în unităţi simbolice şi răspunde la ele. Lucram la reţele neuronale în vremea aceea şi nu am făcut progrese prea mari, aşa că m-am concentrat pe altceva.

În anii '2002 – 2003, m-am gândit din nou la întrebare: ce ar implica un sistem computaţional de cunoştinţe? Munca pe care o

făcusem până atunci mi-ar arătat că greşeam complet în ipoteza mea iniţială despre cum aş putea obţine acest rezultat. Ipoteza mea iniţială fusese că, pentru a putea face un sistem computaţional de cunoştinţe serios, mai întâi trebuie să construieşti un dispozitiv asemănător unui creier, în care să încarci apoi cunoştinţe – la fel cum învaţă oamenii în educaţia standard. Acum mi-am dat seama că nu era o linie clară între ce este inteligent şi ce este pur şi simplu computaţional.

Presupusesem că există un mecanism magic care ne făcea mult mai capabili decât orice este pur computaţional. Dar această presupunere era greşită. Această realizare a condus la Wolfram Alpha. Ce am descoperit era că poţi să iei o colecţie mare de cunoştinţe despre lume şi poţi răspunde automat în baza ei, utilizând doar tehnici computaţionale. Era o metodă alternativă de inginerie – o metodă mult mai analogă cu ceea ce face biologia în evoluţie.

În concluzie, ce faci de obicei când construieşti un program este să-l construieşti pas cu pas. Dar poţi şi să explorezi universul computaţional şi să minezi tehnologia din acest univers. De obicei provocarea este aceeaşi ca în cazul mineritului fizic: adică, să zicem că găseşti un depozit de fier, cobalt sau gadoliniu, cu proprietăţi magnetice speciale, şi foloseşti această capacitate specială în vederea obţinerii unui scop uman, pentru ceva ce vrei tu să facă tehnologia. În cazul materialelor magnetice, există multe căi de a face asta. În ceea ce priveşte programele, aceeaşi poveste. Există tot felul de programe, chiar şi programe mărunte care fac lucruri complicate. Le-am putea pune în slujba unui scop uman util?

Şi cum faci inteligenţele artificiale să-ţi execute obiectivele? Un răspuns ar fi să vorbeşti cu ele în limba naturală a omului. Funcţionează destul de bine când vorbeşti cu Siri. Dar, când vrei să spui ceva mai lung sau mai complicat, nu mai funcţionează la fel de bine. Avem nevoie de un limbaj de programare care poate reprezenta concepte sofisticate într-un mod care poate fi construit progresiv şi nu este posibil în limba naturală. Compania mea a petrecut mult timp încercând să construiască un limbaj bazat pe cunoştinţe care încorporează cunoaşterea lumii direct în limbaj. Abordarea tradiţională de creare a unui limbaj de programare este să faci un limbaj ce reprezintă operaţii pe care computerele ştiu intrinsec cum să le execute: alocarea memoriei, atribuirea unor valori variabilelor, iteraţia, modificarea contoarelor de numere şi aşa mai departe. Esenţialmente, le spui computerelor să facă ceva în condiţiile stabilite de tine. Abordarea mea presupunea crearea unui limbaj care serveşte oamenilor, şi nu computerelor, şi care să preia lucrurile la care se gândeşte omul şi să le convertească într-o formă pe care computerul o poate înţelege. Am putea încapsula cunoştinţele pe care le-am acumulat, atât în ceea ce priveşte ştiinţele, cât şi colectarea datelor,

într-un limbaj pe care l-am putea utiliza ca să comunicăm cu computerele? Aceasta este marea realizare a ultimilor treizeci şi ceva de ani din cariera mea.

În anii 1960, oamenii spuneau lucruri precum „când vom face cutare şi cutare, vom şti că avem inteligenţă artificială”; „când vom rezolva o integrală dintr-un curs de analiză matematică, vom şti că avem inteligenţă artificială”; „când vom putea avea o conversaţie cu un computer şi o vom face să pară umană…” etc. Dificultatea consta în „stai aşa, computerul nu ştie suficiente lucruri despre lume”. Întrebai computerul ce zi din săptămână este şi probabil putea să-ţi răspundă. îl întrebai cine este preşedinte şi probabil nu-ţi putea spune. În acel moment, ştiai că vorbeşti cu un computer, şi nu cu o persoană. Dar acum, când vine vorba despre aceste teste Turing, oamenii care au încercat să conecteze, de exemplu, Wolfram Alpha la boţii lor de testare Turing observă că boţii lor pierd de fiecare dată. Pentru că tot ce trebuie să faci este să începi să-i pui maşinii întrebări sofisticate şi îţi va răspunde la ele! Niciun om nu poate face asta. După câteva întrebări din domenii diferite, niciun om nu va şti răspunsul la toate acele întrebări, însă sistemul va şti. În acest sens, am obţinut deja o inteligenţă artificială bună, la acel nivel.

Apoi mai sunt sarcini care sunt uşoare pentru oameni, dar tradiţional foarte grele pentru maşini. Sarcina standard este identificarea vizuală a obiectelor: ce este acest obiect? Oamenii îl pot recunoaşte şi-l pot descrie simplu, dar computerul nu avea nicio şansă. Însă acum doi ani am scos un mic sistem de identificare a imaginilor şi multe alte companii au făcut ceva similar – al nostru se întâmplă să fie puţin mai bun decât celelalte. Îi arăţi o imagine şi, pentru aproximativ 10.000 de tipuri de lucruri, îţi poate spune ce este. Este distractiv să-i arăţi o pictură abstractă şi să vezi ce spune. Dar face o treabă destul de bună.

Funcţionează cu aceeaşi tehnologie de reţea neuronală pe care şi-au imaginat-o McCulloch şi Pitts în 1943 şi mulţi dintre noi au lucrat la ea la începutul anilor 1980. În anii 1980, oamenii au reuşit să facă OCR – recunoaşterea optică a caracterelor. Au luat cele douăzeci şi şase de litere ale alfabetului şi au zis: Ja să vedem, ăsta este un A? Este un B? Este un C?” şi aşa mai departe. Acest lucru poate fi realizat pentru douăzeci şi şase de posibilităţi diferite, dar nu poate fi făcut pentru 10.000. Problema era cum scalăm întregul sistem care face acest lucru posibil în prezent. Există probabil cinci mii de substantive comune care pot fi transpuse în imagini în limba engleză, zece mii dacă includem şi plante speciale şi insecte pe care oamenii le-ar recunoaşte cu o anumită frecvenţă. Ce am făcut a fost să ne antrenăm sistemul cu treizeci de milioane de imagini care reprezintă toate aceste obiecte. Este o reţea neuronală mare, complicată şi haotică.

Probabil că detaliile reţelei nu contează, dar sunt necesare un cvadrilion de operaţii ale unităţii de procesare grafică pentru această instruire.

Sistemul nostru este impresionant, pentru că aproape că poate să facă ce face un om. Are aproximativ acelaşi volum de informaţii pe care-l folosesc şi oamenii – un copil vede aproximativ acelaşi număr de imagini în primii doi ani ai vieţii sale. Aproximativ acelaşi număr de operaţii trebuie făcute şi în procesul de învăţare, utilizând aproximativ acelaşi număr de neuroni, cel puţin în primul nivel al cortexului nostru vizual. Detaliile sunt diferite; modul în care funcţionează aceşti neuroni artificiali nu prea are de-a face cu modul în care funcţionează neuronii creierului. Dar conceptul este similar şi există o anumită universalitate pentru ce se întâmplă. La nivel matematic, este o compoziţie cu foarte multe funcţii, cu anumite proprietăţi de continuitate care permit utilizarea anumitor metode de calcul ce permit instruirea incrementală a sistemului. Date fiind aceste atribute, poţi obţine ceva ce face ce fac şi creierele omeneşti în recunoaşterea psihologică.

Dar constituie asta inteligenţă artificială? Sunt câteva componente de bază. Este vorba despre recunoaşterea psihologică, transformarea vocii în text, traducerea limbii – lucruri pe care oamenii reuşesc să le facă cu diverse grade de dificultate. În principiu, acestea sunt câteva dintre elementele prin care facem maşinile să pară omeneşti în ceea ce fac. Pentru mine, unul dintre lucrurile interesante a fost să încorporez aceste capacităţi într-un limbaj precis şi simbolic pentru a reprezenta lumea de zi cu zi. Acum avem un sistem care poate spune „acesta este un pahar cu apă”. Putem trece de la imaginea unui pahar cu apă la conceptul paharului cu apă. Acum trebuie să inventăm un limbaj simbolic efectiv pentru a reprezenta aceste concepte.

Am început prin a încerca să reprezint matematic cunoştinţe tehnice şi am continuat cu alte tipuri de cunoştinţe. Am făcut o treabă foarte bună cu reprezentarea cunoştinţelor obiective din lume. Acum problema este reprezentarea discursului uman de zi cu zi într-un mod precis şi simbolic – un limbaj bazat pe cunoştinţe care are ca scop comunicarea între oameni şi maşini, astfel încât oamenii să-l poată citi şi maşinile să-l poată înţelege. De exemplu, putem spune „x este mai mare decât 5”. Acesta este un predicat. Ai putea spune şi „vreau o bucată de ciocolată”. Şi acesta este un predicat. Are un „vreau” în el. Trebuie să găsim o reprezentare simbolică precisă pentru dorinţele pe care le exprimăm în limba naturală omenească.

La sfârşitul anilor 1600, Gottfried Leibniz, John Wilkins şi alţii se preocupau de ceea ce numeau ei limbajele filosofice, adică reprezentări complete, universale şi simbolice ale lucrurilor din lume. Poţi analiza limbajul filosofic al lui John Wilkins şi poţi vedea cum a

împărţit ce era important în lume la vremea aceea. Unele aspecte ale condiţiei umane au fost aceleaşi încă din anii 1600. Altele sunt foarte diferite. Secţiunea sa despre moarte şi diversele forme ale suferinţei umane era imensă; în ontologia de astăzi este mult mai mică. Este interesant să vezi cum un limbaj filosofic de astăzi diferă de un limbaj filosofic de la jumătatea secolului al XVII-lea. Este o măsură a progresului nostru. Multe astfel de încercări de formalizare au avut loc de-a lungul anilor. În matematică, de exemplu, *Principia Mathematica* a lui Whitehead şi Russell din 1910 a reprezentat cea mai mare paradă de erudiţie. Au existat şi încercări înaintea lor din partea lui Gottlob Frege şi Giuseppe Peano, care au fost puţin mai modeste în prezentare. În cele din urmă, s-au înşelat în ceea ce credeau ei că ar trebui formalizat: considerau că trebuie să formalizeze un proces de dovezi matematice, care se pare că nu este lucrul de care sunt interesaţi cei mai mulţi oameni.

Cu privire la un corespondent modern al testului Turing, este o întrebare interesantă. Există încă botul conversaţional, care este ideea lui Turing. Acesta nu a fost rezolvat încă. Va fi rezolvat – singura întrebare este „care este aplicaţia pentru care este rezolvat?" Mult timp aş fi întrebat „de ce să ne pese?" – pentru că am crezut că aplicaţia principală ar fi în domeniul asistenţei clienţilor, care nu era prea sus pe lista mea. Dar asistenţa clienţilor, unde încerci să creezi o interfaţă, este fix locul în care ai nevoie de acest limbaj conversaţional.

O diferenţă mare între vremea lixi Turing şi a noastră este metoda de comunicare cu computerele. Pe vremea lui, introduceai ceva în maşină şi maşina îţi afişa un răspuns pe ecran. În ziua de astăzi, răspunde cu un ecran – de exemplu, atunci când vrei să cumperi un bilet la film. Cum este diferită tranzacţia făcută cu o maşină faţă de tranzacţia făcută cu un om? Răspunsul principal este că există un afişaj vizual. Te întreabă ceva şi tu apeşi un buton, putând vedea rezultatul imediat. De exemplu, în Wolfram Alpha, când este utilizat de Siri, dacă răspunsul este scurt, Siri îţi va spune răspunsul scurt. Dar cei mai mulţi oameni vor afişajul vizual, care arată informaţiile grafice pentru cutare şi cutare. Aceasta este o formă de comunicare nonumană care se dovedeşte a fi mai bogată decât comunicarea umană tradiţională rostită sau tastată. În majoritatea comunicărilor de la om la om, suntem limitaţi la limbajul pur, pe când în comunicarea dintre computer şi om avem acest canal de comunicare mult mai larg – comunicarea vizuală.

Multe dintre cele mai puternice aplicaţii ale testului Turing nu mai sunt relevante acum, pentru că avem acest canal suplimentar de comunicare. De exemplu, iată unul pe care-l urmărim în prezent. Este un bot care comunică pe tema scrierii de cod: îi spui „vreau să scriu

un program. Vreau să facă asta". Botul va spune „am scris această bucată de cod. Face asta. Este ceea ce ai vrut?" Bla, bla, bla. Este un bot cu du-te-vino. Dezvoltarea unor astfel de sisteme este o problemă interesantă, pentru că trebuie să aibă un model pentru om dacă încearcă să-ţi explice ceva. Trebuie să ştie ce nu înţelege omul.

Ce mi-a fost mie greu să înţeleg este: care este scopul unui test Turing? Care este motivaţia? Ca jucărie, se poate face un mic bot de chat cu care oamenii să discute. Aceasta ar fi următoarea mare descoperire. Runda curentă de învăţare profundă – în special reţelele neuronale recurente – face modele destul de bune de vorbire şi scriere umane. Putem să tastăm, de exemplu: „Cum te simţi astăzi?" şi ştie aproape tot timpul ce răspuns să-ţi dea. Dar vreau să-mi dau seama dacă pot să-l fac să răspundă automat la e-mailurile pe care le primesc. Ştiu că răspunsul este nu. Pentru mine, un test Turing bun va fi atunci când un bot ar putea să răspundă la majoritatea e-mailurilor pe care le primesc. Acesta este un test dificil. Va trebui să înveţe aceste răspunsuri de la omul la care este conectat e-mailul. Aş putea avea un mic avantaj, pentru că am colectat informaţii despre mine în ultimii douăzeci şi cinci de ani. Am toate e-mailurile primite în ultimii douăzeci şi cinci de ani, istoricul tuturor tastelor apăsate de douăzeci de ani. Ar trebui să pot să educ un avatar, o inteligenţă artificială, care să facă ce pot face eu – poate chiar mai bine decât mine.

Oamenii îşi fac griji gândindu-se la scenariul în care inteligenţele artificiale preiau controlul. Cred că înainte de asta se va întâmpla ceva mult mai amuzant, într-un anumit sens. Inteligenţa artificială va şti ce intenţionezi şi se va pricepe să-şi dea seama cum să obţină ce doreşti. Îi spun GPS-ului maşinii mele că vreau să ajung la o anumită destinaţie. Nu ştiu unde naiba mă aflu, doar îmi urmez GPS-ul. Copiilor mei le place să-mi aducă aminte că, odată, când aveam un GPS foarte vechi – tipul care îţi spunea „Faceţi prima la stânga, prima la dreapta etc". –, am ajuns pe unul dintre pontoanele care coborau în portul din Boston.

Mai important este faptul că va exista o inteligenţă artificială care îţi cunoaşte istoricul şi ştie că, atunci când comanzi cina online, probabil vei vrea aia şi aia, sau atunci când îi trimiţi un e-mail unei anumite persoane, probabil vei vorbi cu ea despre cutare şi cutare. Din ce în ce mai pregnant, inteligenţele artificiale ne vor sugera ce ar trebui să facem şi bănuiesc că, de cele mai multe ori, oamenii vor fi de acord. Este un sfat bun – mai bine decât la ce te-ai fi gândit tu.

Cât priveşte scenariul în care acestea preiau controlul, poţi face rău cu tehnologia şi poţi face bine cu ea. Unii oameni vor încerca să facă lucruri teribile cu ea, iar alţii, lucruri bune. Unul dintre lucrurile care-mi plac la tehnologia de astăzi este egalizarea pe care a produs-o.

Obişnuiam să mă mândresc cu faptul că am un computer mai bun decât toate cunoştinţele mele; acum avem cu toţii aceleaşi computere. Avem aceleaşi telefoane şi aproximativ aceeaşi tehnologie poate fi folosită de un procent important dintre cei 7 miliarde de oameni de pe planetă. Tehnologia regelui nu mai este diferită de a tuturor celorlalţi. Acesta este un avans important.

Marea frontieră acum cinci sute de ani era alfabetizarea. Astăzi este programarea. Programarea aşa cum se face în prezent va deveni depăşită în scurt timp. De exemplu, oamenii nu mai învaţă limbaje de programare, deoarece computerele se pricep mai bine să scrie programe decât oamenii şi doar un număr mic de oameni trebuie să cunoască detaliile despre cum limbajul comun este compilat în limbajele de programare. Foarte mult din ceea ce fac armate de programatori din zilele noastre este plictisitor de similar. Nu mai există niciun motiv pertinent pentru care oamenii să scrie cod Java sau Java-Script. Putem automatiza procesul de programare astfel încât ce este important merge de la ce vrea omul la execuţia de către maşină a dorinţei sale, cât mai automatizat cu putinţă. Astfel este crescută egalizarea, ceea ce mă interesează pe mine. În trecut, dacă doreai să scrii o bucată serioasă de cod sau program pentru ceva important şi real, era mult de muncă. Trebuia să ştii multe despre programare, trebuia să investeşti luni de zile, trebuia să angajezi programatori care ştiau ce fac sau trebuia să înveţi singur. Era o investiţie mare.

În prezent, nu mai este cazul. O piesă de cod de o linie face deja ceva interesant şi util. Permite unui număr mare de oameni care nu puteau să-şi determine computerele să facă ceva pentru ei acum să o poată face. Aş vrea să văd mulţi copii din jurul lumii învăţând noile capacităţi ale programării bazate pe cunoştinţe şi apoi să producă coduri care sunt efectiv la fel de sofisticate asemenea celor produse de experţi. Acest lucru este realizabil. Suntem în punctul în care oricine poate învăţa să facă programare bazată pe cunoştinţe şi, mai important, poate învăţa să gândească computaţional.

Mecanica efectivă a programării este uşoară acum. Ce este dificil este să-ţi imaginezi ceva într-o manieră computaţională.

Cum predai gândirea computaţională? În ceea ce priveşte programarea, este o întrebare interesantă. Să ne gândim la nanotehnologie. Cum am obţinut nanotehnologia? Am luat tehnologia aşa cum o înţelegem la scară mare şi am făcut-o foarte mică. Cum aducem un cip CPU la scară atomică? Fundamental, folosim aceeaşi arhitectură ca în cazul cipului CPU pe care-l ştim şi-l iubim cu toţii. Aceasta nu este singura abordare pe care o avem. Observând ce pot face programele simple, este posibil să transformăm componente simple şi învechite şi, cu compilatorul potrivit, să facem din ele ceva

interesant. Nu facem calcule la scară moleculară încă, pentru că tehnologia de care dispunem ar putea să producă ce dorim în zece ani, probabil. Dar avem componentele necesare pentru a construi un computer universal. Poate nu ştim cum să programăm cu aceste componente, dar, căutând în spaţiul programelor posibile, vom începe să colectăm modulele, apoi putem crea un compilator pentru ele. Surprinzător este că lucrurile vechi pot fi transformate în ceva sofisticat, iar etapa de compilare nu este atât de înfiorătoare pe cât credem.

Simpla căutare prin universul computaţional şi încercarea de a găsi programe – module – care sunt interesante este o abordare bună. O abordare inginerească mai tradiţională – încercarea de a construi un computer universal doar prin gândire pură – este o treabă dificilă. Asta nu înseamnă că nu se poate face, dar eu cred că vom putea face lucruri minunate doar căutând componentele şi programele posibile pe care le putem face cu ele. Apoi ne întoarcem la problema conectării scopurilor omeneşti la ce ne oferă sistemul.

O întrebare al cărei răspuns mă interesează este: „Cum va arăta lumea când majoritatea oamenilor vor putea să scrie cod?" Am avut o tranziţie, poate acum cinci sute de ani, când doar scribii şi un procent mic din populaţie puteau să citească şi să scrie limba naturală. Astăzi un procent mic din populaţie poate să scrie cod. Cea mai mare parte din codul scris este doar pentru computere.

Nu înţelegi lucrurile dacă citeşti codul. Dar va veni vremea când, ca urmare a lucrurilor pe care am încercat eu să le fac, codul va fi la un nivel suficient de înalt încât să ofere o descriere minimă pentru ceea ce încerci să faci. Va fi o bucată de cod care poate fi înţeleasă de oameni, dar poate fi şi executată de maşini.

Codarea este o formă de exprimare, la fel cum a scrie într-o limbă naturală este o formă de exprimare. Pentru mine, bucăţile simple de cod sunt poetice – exprimă ideile foarte curat. Există o latură estetică, la fel cum există şi în exprimarea într-o limbă naturală. O caracteristică a codului este că poate fi executat imediat; nu este ca scrisul. Când scrii ceva, cineva trebuie să citească ce ai scris şi creierul care citeşte trebuie să absoarbă gândurile care vin de la persoana care a scris. Priveşte felul în care au fost transmise cunoştinţele în istoria lumii. La nivelul zero, o formă de transmitere a cunoştinţelor este genetica – adică există un organism şi progeniturile sale au aceleaşi caracteristici pe care le are el. Apoi urmează transmiterea de cunoştinţe, care se întâmplă cu lucruri precum recunoaşterea psihologică. O creatură nou-născută are o reţea neuronală cu câteva conexiuni aleatorii în ea şi, pe măsură ce creatura se mişcă prin lume, începe să recunoască tipuri de obiecte şi învaţă acele cunoştinţe.

Apoi urmează nivelul care a reprezentat marea realizare a speciei

noastre, care este limba naturală: abilitatea de a reprezenta cunoştinţe suficient de abstract încât să le putem comunica de la creier la creier. Este posibil ca limba naturală să fie una dintre cele mai importante invenţii ale speciei noastre. Este ceea ce ne-a condus, în multe privinţe, la civilizaţia pe care o avem.

Mai există un nivel şi, probabil, într-o bună zi, va avea un nume mai interesant. Cu programarea bazată pe cunoştinţe, avem o modalitate de a crea o reprezentare efectivă a lucrurilor reale din lume într-un mod precis şi simbolic. Nu doar că poate fi înţeleasă de creiere şi poate fi comunicată altor creiere şi computere, dar mai este şi imediat executabilă.

Limba naturală ne-a dat civilizaţia, dar programarea bazată pe cunoştinţe ce ne va da? Un răspuns greşit este că ne va da civilizaţia inteligenţelor artificiale. Asta nu vrem să se întâmple, pentru că inteligenţele artificiale vor comunica foarte bine una cu cealaltă şi noi vom fi scoşi din schemă, pentru că nu există nicio limbă intermediară, nicio interfaţă cu creierele noastre. Unde ne va conduce acest al patrulea nivel de comunicare bazat pe cunoştinţe? Dacă ai fi fost Zog în peşteră şi ţi-ai fi dat seama că începem să folosim limbajul, ţi-ai fi putut imagina civilizaţia în care trăim? Ce ar trebui să ne imaginăm acum?

Această întrebare se leagă de întrebarea cum ar arăta lumea dacă majoritatea oamenilor ar putea să codeze. În mod clar, multe lucruri triviale s-ar schimba, contractele s-ar scrie în cod, reţetele de restaurant s-ar scrie în cod şi aşa mai departe. Lucruri simple ca acestea s-ar schimba. Dar s-ar schimba şi lucruri mult mai profunde. Creşterea nivelului de alfabetizare ne-a dat birocraţia, de exemplu, care exista deja, dar a fost accelerată masiv, conferindu-ne o profunzime mai mare pentru sistemele guvernamentale, în bine şi în rău. Cum se leagă lumea codului de lumea culturală?

Să ne gândim la educaţia din licee. Dacă avem gândire computaţională, cum ne afectează ea modul în care studiem istoria? Cum ne afectează modul în care studiem limbile, ştiinţele sociale şi altele? Răspunsul este că are un efect puternic. Imaginează-ţi că scrii un eseu. Astăzi, materia primă pentru eseul unui licean tipic este ceva ce a fost scris deja; elevii nu pot, de obicei, să genereze cunoştinţe noi cu uşurinţă. Dar în lumea computaţională nu va mai fi cazul. Dacă elevii ştiu ceva despre cum se scrie un cod, vor accesa toate informaţiile istorice digitalizate şi vor ajunge să creeze ceva nou. Apoi vor scrie un eseu despre ceva ce au descoperit. Realizarea programării pe bază de cunoştinţe se bazează pe faptul că nu mai este sterilă, pentru că are cunoştinţele lumii întreţesute în limba pe care o utilizezi pentru a scrie codul.

Computaţii sunt în tot universul: într-un fluid turbulent care produce un tipar complicat de curgere, în mecanica celestă a interacţiunilor planetare, în creiere. Dar are tehnica de calcul un scop? Poţi pune această întrebare referitoare la orice sistem? Vremea are un scop? Clima are un scop?

Cineva care priveşte Pământul din spaţiu poate să-şi dea seama dacă ceva are scop acolo jos? Există o civilizaţie acolo? În Marele Lac Sărat din Utah se află o linie dreaptă. Se pare că este un dig circulabil care împarte două zone ale lacului ce au alge de culori diferite, deci este o linie dreaptă foarte dramatică. Există un drum în Australia care este lung şi drept. Există o cale ferată în Siberia care este lungă şi se aprind lumini când trenul opreşte într-o gară. Deci, din spaţiu, se pot vedea linii drepte şi tipare.

Dar sunt acestea exemple suficient de clare de scopuri clare pe Pământ dacă sunt observate din spaţiu? De fapt, cum ne-am da seama dacă în spaţiu sunt extratereştri? Cum ne dăm seama dacă un semnal pe care-l interceptăm indică un scop? Pulsarii au fost descoperiţi în 1967, când am detectat o oscilaţie periodică la fiecare secundă. Prima întrebare a fost dacă nu cumva este un semnal? întrucât ce altceva ar emite un semnal periodic? S-a dovedit a fi o stea neutronică rotativă.

Un criteriu de aplicat unui fenomen care prezintă potenţialul de a avea un scop este dacă îndeplineşte condiţiile acţiunii minime pentru atingerea scopului. Dar înseamnă asta că a fost construit pentru scopul respectiv? Mingea se rostogoleşte în josul pantei datorită atracţiei gravitaţionale. Sau mingea se rostogoleşte în josul pantei pentru a satisface principiul acţiunii minime. De obicei, acestea sunt cele două explicaţii pentru o acţiune care pare că are un scop: explicaţia mecanicistă şi explicaţia teleologică. Practic, toată tehnologia noastră existentă eşuează la testul acţiunii minime în atingerea scopului său. Cele mai multe lucruri pe care le construim noi sunt înrădăcinate adânc în istoricul tehnologic şi sunt incredibil de nonminime în ceea ce priveşte atingerea scopului. Gândeşte-te la un cip CPU; nu-i nicio şansă ca aceea să fie calea minimă de a obţine ce obţine un cip CPU.

Această problemă a identificării discernământului este una grea. Este o întrebare importantă, pentru că zgomotele radio venite din galaxie sunt foarte similare transmisiunilor CDMA ale telefoanelor mobile. Aceste transmisii folosesc secvenţe de pseudozgomot, care se întâmplă să aibă anumite proprietăţi de repetabilitate. Dar sunt percepute ca zgomot şi sunt concepute ca zgomot pentru a nu interfera cu alte canale. Problema devine şi mai neclară. Dacă am analiza o secvenţă de numere prime generate de un pulsar, ne-am întreba ce le-a generat. Ar însemna că o întreagă civilizaţie a crescut, a descoperit numerele prime şi a inventat computerele şi transmiţătoarele radio, după care a făcut această emisie? Sau este

vorba doar despre un proces fizic care emite numere prime? Există un automat celular care emite numere prime. Poţi vedea cum funcţionează dacă-l dezasamblezi. Are ceva mic ce ricoşează în el şi din acest mecanism reiese o secvenţă de numere prime. Nu a avut nevoie de întreaga istorie a civilizaţiei şi a biologiei pentru a ajunge în acest punct.

Nu cred că există un „scop" abstract, se Nu cred că există sensuri abstracte. Universul are un scop? Cu această întrebare eşuezi în teologie. Nu există niciun sens important în care să existe o noţiune abstractă a scopului. Scopul este ceva ce reiese din istorie.

Unul dintre lucrurile care ar putea fi adevărate despre lumea noastră este că poate trecem prin toată istoria, biologia şi civilizaţia noastră şi, după toate acestea, răspunsul este „42" sau ceva de genul acesta. Am trecut prin patru miliarde de ani de diverse tipuri de evoluţie şi rezultatul este „42".

Nu se va întâmpla nimic de acest gen, datorită ireductibilităţii computaţionale. Există procese computaţionale pe care le putem parcurge în care nu există nicio scurtătură pentru proces. O mare parte din ştiinţă se preocupă de găsirea unei scurtături pentru computaţiile făcute de natură. De exemplu, dacă facem mecanică celestă şi vrem să prezicem poziţia planetelor de peste un milion de ani, am putea urma ecuaţiile pas cu pas. Dar marea realizare a ştiinţei este că putem să scurtăm acest proces şi să reducem calculele. Putem fi mai inteligenţi decât Universul şi-i putem prezice punctul final fără a trece prin toate etapele. Dar chiar şi cu o maşină suficient de inteligentă şi cu o matematică suficient de bună, nu putem ajunge la punctul final fără a trece prin etape. Unele detalii sunt ireductibile. Trebuie să urmăm ireductibil acele etape. Acesta este motivul pentru care istoria înseamnă ceva. Dacă am putea ajunge la punctul final fără a trece prin etape, istoria ar fi, într-un anumit sens, inutilă.

Aşadar ideea nu este că noi suntem inteligenţi şi toate celelalte din lume nu sunt. Nu este nicio diferenţă abstractă enormă între noi şi nori, şi automatele celulare. Nu putem spune că această reţea neuronală asemănătoare creierului este calitativ diferită de acest sistem automat al celularului. Diferenţa este o diferenţă detaliată. Această reţea neuronală asemănătoare creierului a fost produsă de istoria lungă a civilizaţiei, în timp ce automatul celular a fost creat de computerul meu în ultima microsecundă.

Problema inteligenţei artificiale abstracte este similară problemei recunoaşterii inteligenţei extraterestre: cum stabileşti dacă are sau nu un scop? Aceasta este o întrebare la care eu consider că nu s-a dat un răspuns. Spunem ceva de genul „Ei bine, inteligenţa artificială va fi inteligentă când va putea face cutare şi cutare". Când va putea găsi numere prime. Când va putea produce cutare şi cutare obiect. Dar

există multe alte căi de a obţine acele rezultate. Din nou, nu există nicio linie despărţitoare clară între inteligenţă şi simpla tehnică de calcul.

Este o altă parte a poveştii copernicane: obişnuiam să credem că Pământul este centrul Universului. Acum credem că suntem speciali pentru că avem inteligenţă şi nimic altceva nu o mai are. Mă tem că vestea proastă ar putea fi că aceasta nu este o distincţie.

Iată unul dintre scenariile mele. Să zicem că va veni o vreme în care conştiinţa umană se va putea încărca uşor pe o platformă digitală, virtualizată şi cu toate cele, şi, destul de repede, am avea o cutie cu un trilion de suflete. În cutie se află un trilion de suflete, toate virtualizate. În cutie, computaţiile neîntrerupte vor fi moleculare – poate derivate din biologie, poate nu. Dar cutia va face tot soiul de lucruri complicate. Lângă cutie se află o piatră. În piatră, se întâmplă tot timpul lucruri complicate, tot felul de particule subatomice care fac tot felul de lucruri. Care este diferenţa dintre piatră şi cutia cu un trilion de suflete? Răspunsul este că detaliile legate de ceea ce se întâmplă în cutie au fost derivate dintr-o istorie lungă a civilizaţiei umane, inclusiv ce au văzut oamenii pe YouTube cu o zi înainte. În timp ce piatra are istoria ei geologică lungă, dar nu şi istoria civilizaţiei noastre.

Realizarea că nu există o distincţie veritabilă între inteligenţă şi simpla tehnică de calcul te face să-ţi imaginezi acest viitor – punctul final al civilizaţiei noastre ca o cutie cu un trilion de suflete, fiecare dintre ele jucând, practic, un joc video, pentru totdeauna. Care este „scopul” aici?

[←1]

Pe parcursul cărţii, vom folosi titlul cărţii în limba engleză, deoarece această carte nu a fost tradusă în limba română. O traducere aproximativă a titlului ar fi *Folosirea omului de către om* (n.t.).

[←2]

In original, Thirteen Ways of Looking at a Blackbird (n.t.).

[←3]

In original, Cybernetics: or Control and Communication in the Animal and the Machine (n.t.).

[←4]

i.e. american (n.t.).

[←5]

Demonul lui Maxwell este o fiinţă imaginară, inteligentă, de dimensiuni moleculare, care îşi permite să încalce principiul al doilea al termodinamicii (n.t.).

[←6]

Trecerea unui bit de la 0 la 1 sau de la 1 la 0 (n.t.).

[←7]

În original, 7he Singularity IS Near **(n.t.).**

[←8]

în original, „robots falling down“ (n.t.).

[←9]

în original, „singulariteers“ (n.t.).

[←10]

Omohundro, „The Basic AI Drives“, in Proceedings of the First AGI Conference, 171; and in P. Wang, B. Goertzel, and S. Franklin, edArtificial General Intelligence (Amsterdam, The Netherlands: IOS Press, 2008).

[←11]

Raportul AI 100 (Peter Stone et al.), sponsorizat de Universitatea Stanford, include următoarele: „Spre deosebire de filme, nu se întrezăreşte nicio rasă de roboţi supraumani la orizont şi, probabil, nici nu este posibil“, https://ail00. stanford.edu/2016-report (n.a.).

[←12]

Elon Musk, Stephen Hawking şi alţii (printre care, se pare, şi autorul) au primit în 2015 premiul Ludditul Anului de la Fundaţia de Inovaţie în Tehnologia Informaţiei: https://itif.org/publications/2016/01/19/artificial-in- telligence-alarmists-win-itif%E2%80%99s-annual-luddite-award (n.a.).

[←13]

Rodney Brooks, de exemplu, spune că este imposibil ca un program să fie „suficient de inteligent astfel încât să poată inventa moduri de a submina societatea umană pentru a atinge scopuri stabilite pentru el de către oameni, fără a înţelege modurile în care cauzează probleme pentru aceiaşi oameni“, http://rodneybrooks.com/the-seven-deadly-sins-of-predicting-the-future- of-ai (n.a.).

[←14]

In original, „cooperative inverse-reinforcement learning“ (n.t.).

[←15]

Consultaţi Hadfield-Menell et al., ,,The Off-Switch Game", https://arxiv.
org/pdf/1611.08219.pdf (n.a.).

[←16]

The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), p. 96 (n.a.).

[←17]

Joseph Levine, JMaterialism and Qualia: The Explanatory Gap“, Pacific Philosophical Quarterly 64 (1983): 354-61 (n.a.).

[←18]

The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), p. 181 (n.a.).

[←19]

The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), p. 181 (n.a.).

[←20]

în cartea Darwins Dangerous Idea, 1995, p. 109, am definit versiunea capitalizată, „vast“ însemnând mult mai mult decât astronomic, şi noţiunea complementară „efemer“ pentru a înlocui obişnuitele exagerări infinit şi infinitezimal pentru discutarea acelor posibilităţi care nu sunt oficial infinite, însă infinite în toate scopurile practice (n.a.).

[←21]

Aylin Caliskan-Islam, Joanna J. Bryson $i Arvind Narayanan, „Semantics Derived Automatically from Language Corpora Contain Human- Like Biases“, Science 356, nr. 6334 (14 aprilie, 2017): 183-86, DOL10.1126/ science.aal4230 (n.a.).

[←22]

Joanna J. Bryson, „Robots Should Be Slaves“, în Close Engagement with Artificial Companions, Yorick Wilks, ed. (Amsterdam, The Netherlands: John Benjamins, 2010), 63- 74; http:// www.cs.bath.ac.uk/~jjb/ftp/Bryson- Slaves-Joanna J. Bryson, „Patiency Is Not a Virtue: AI and the Design of Ethical Systems“, https://www.cs.bath.ac.uk/~jjb/ftp/Bryson-Patiency- AAAISS16.pdf (n.a.).

[←23]

În original, "On Computable Numbers with an Application to the Entscheidungsproblem" (n.t.).

[←24]

„A Structure for Deoxyribose Nucleic Acid“, Nature 171 (1953): 737-38 (n.a.).

[←25]

J. von Neumann, „First Draft of a Report on the EDVAC“, IEEE Annals of the History of Computing 15 (1993): 27- 75. Von Neumann este listat ca fiind singurul autor, în timp ce alţii au contribuit la conceptele pe care le-a prezentat el; astfel întreaga arhitectură i-a fost atribuită lui (n.a.).

[←26]

În original, "buffer overrun". Microsoft Language Portal oferă şi varianta
de traducere "depăşire a tamponului" (n.t.).

[←27]

Science 177, nr.4047 (4 august, 1972) 393-96 (n.a.)

[←28]

„Experiment mental” în limba germană (n.t.).

[←29]

Vincent C. Miiller and Nick Bostrom, "Future Progress in Artificial Intelligence: A Survey of Expert Opinion", în Fundamental Issues of Artificial lntelligence, ed. Vincent C. Miiller (Elveţia: Springer lnternational Publishing, 2016), 555- 72, https://nickbostrom.com/papers/survey.pdf (n.a.).

[←30]

Upton Sinclair, L Candidate far Governor: And How I Got Licked (Berkeley,CA: University of California Press, 1994), 109 (n.a.).

[←31] https://futureoflife.org/ai-principles/ (n.a.).

[←32]

Vezi, de exemplu, Hannah Arendt, Eichmann in Jerusalem: A Report on the Banality of Evil (New York: Penguin Classics, 2006) (n.a.).

[←33]

Vezi Elizabeth Kolbert, The Sixth Extinction: An Unnatural History (New York; Henry Holt, 2014) (n.a.).

[←34]

In original, J am the master of my fate, /1 am the captain of my soul“, traducere de Neculae Leonard (n.t.)

[←35]

Retipărită post-mortem în Philosophia Mathematica 4, nr. 3 (1966): 256-60 (n.a.).

[←36]

Irving John Good, „Speculations Concerning the First Ultraintelligent Machine“, Advances in Computers 6 (Cambridge, Massachusetts: Academic Press, 1965): 31-88 (n.a.).

[←37]
Katja Grace et al., „When Will AI Exceed Human Performance? Evidence from AI Experts“, https://arxiv.org/pdfrl705.08807.pdf (n.a.).

[←38]

Neil Gross şi Solon Simmons, „The Social and Political Views of American College and University Professors“, în Professors and Their Politicsy ed. N. Gross şi S. Simmons (Baltimore: Johns Hopkins University Press, 2014) (n.a.).

[←39]

Termen folosit de Mihnea Gafiţa în traducerea romanului 1984, de George Orwell (n.t.).

[←40]

Steven Pinker, „Safety“, Enlightenment Now: The Case for Reason, Science, Humanismy and Progress (New York: Penguin, 2018). (n.a.)

[←41]

Traducere Ion Vinea (n.t.).

[←42]

Imitarea anumitor comportamente fără a le înţelege (în original „aping“) foloseşte trucuri înnăscute, cum ar fi sistemul neuronilor în oglindă. Dar comportamentele imitate în acest fel sunt foarte limitate în ceea ce priveşte complexitatea. Vezi Richard Byrne, „Imitation as Behaviour Parsing“, Philosophical Transactions of the Royal Society B 358, nr. 1431 (2003): 529-36 (n.a.).

[←43]
Karl Popper, Conjectures and Refutations (Abingdon, Oxon: Roudedge, 1963) (n.a.).

[←44]

Matt Ridley, în cartea *The Raţional Optimist*, subliniază pe bună dreptate efectul pozitiv al populaţiei asupra ratei de progres. Dar acesta nu a fost niciodată cel mai mare factor: să ne gândim, de exemplu, la Atena antică faţă de restul lumii la momentul respectiv (n.a.).

[←45]

Alfred, Lord Tennyson, The Revenge (1878) (n.a.).

[←46]

Acronim de la „Totally Obedient Moron“ (n.t.).

[←47]

Acronim de la „Disobedient Autonomous Thinking Application“ (n.t.).

[←48]

In original, „inverse-reinforcement leaming“ (n.t.).

[←49]

In original, „bounded optimality“ (n.t.).

[←50]

DIY, acronim pentru Do It Yourself folosit des în limba engleză, ce exprimă ideea de a face ceva pe cont propriu, fără ajutorul profesioniştilor (n.t.).

[←51]

De la Quant, acronim pentru Quantitative Analysis for Economics and Finance - analiză cantitativă pentru economie şi finanţe. Termen folosit şi cu sensul de investitori specializaţi de pe Wall Street care-i ajută pe moguli să câştige şi mai mulţi bani (n.t.).

[←52]

Teorie conspiraţionistă şi pseudoştiinţifică, respinsă de majoritatea oamenilor de ştiinţă (n.t.).

[←53]

Norbert Wiener, „A Scientist Rebels“, Atlantic Monthly, ianuarie 1947 (n.a.).

[←54]

Warren Weaver, „Recent Contributions to the Mathematical Theory of Communication“, în Claude Shannon şi Warren Weaver, The Mathematical Theory of Communication (Urbana, IL: University of Illinois Press, 1949), 8 (text evidenţiat în original). Lucrările lui Shannon din 1948 au fost republicate în acelaşi volum (n.a.).

[←55]
Matthew Arnold, Culture and Anarchy, ed. Jane Garnett (Oxford, UK: Oxford University Press, 2006) (n.a.).

[←56]

The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), 17-18 (n.a.).

[←57]
https://strikemag.org/bullshit-jobs (n.a.).

[←58]

Vezi, de exemplu, cartea lui Dennett, From Bacteria to Bach and Back: 'Ihe Evolution of Minds (New York: W.W. Norton, 2017) (n.a.).

[←59]
https://givingpledge.org/About. aspx (n.a.).

[←60]
Brenden M. Lake, Ruslan Salakhutdinov si Joshua B. Tenenbaum, „Human-Level Concept Learning Through Probabilistic Program Induction“, revista Science 350, nr. 6266 (2015): 1332-38 (n.a.).

[←61]

„Blicket“ este un concept introdus de Nancy Soja ce presupune un tip de obiect nou, cu anumite proprietăţi ce pot fi descoperite de om prin anumite experimente legate de cauzalitate sau percepţie (n.t.).

[←62]

A. Gopnik, T. Griffiths şi C. Lucas, „When Younger Learners Can Be Better (or at Least More Open-Minded) Than Older Ones“, Current Directions in Psychological Science 24, nr. 2 (2015): 87-92 (n.a.).

[←63]

A. Gopnik çi colaboratorii, „Changes in Cognitive Flexibility and Hypothesis Search Across Human Life History from Childhood to Adolescence to Adulthood“, PNAS114, nr. 30 (2017): 7892-99 (n.a.).

[←64]

L. Schulz, „The Origins of Inquiry: Inductive Inference and Exploration in Early Childhood“, Trends in Cognitive Sciences 16, nr. 7 (2012): 382-89 (n.a.).

[←65]

A. Gopnik, The Gardener and the Carpenter (New York: Farrar, Straus 6c Giroux, 2016), capitolele 4 $i 5 (n.a.)

[←66]

William M. Grove and Paul E. Meehl, "Comparative Efficiency of Informal (Subjective, lmpressionistic) and Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical- Statistical Controversy", Psychology, Public Policy, and Law 2, no. 2 (1996): 293- 323 (n.a.).

[←67]

TED Talk,January 2014, https:// www.ted.com/ speakers/ anne_ milgram(n.a.).

[←68]

„Then, Doctors «All Anxious» About Test- tube Baby“, http://edition.cnn. com/2003/HEALTH/parenting/07/25/cnna.copperman (n.a.).

[←69] https://www.nbcnews.com/news/us-news/lawyer-denying-chimpanzees- rights-could-backfire-disabled-n734566 (n.a.).

[←70]

în limba engleză, „left“ are sensul de „stânga“, dar şi de „lăsat“ sau „rămas“ (n.t.).

[←71]

The Human Use of Human Beings (Boston: Houghton Mifflin, 1954), 46 (n.a.).

[←72]

Narration in Stephen Hawkings The Meaning of Life (Smithson Productions, Discovery Channel, 2012) (n.a.).

[←73]

Mary Catherine Bateson, 1999 foreword to Gregory Bateson, Steps to an Ecology of Mind (Chicago: University of Chicago Press, 1972), xi (n.a.).

[←74]

Steps to an Ecology of Mind, 452 (n.a.).

[←75]

Steps to an Ecology of Mind, 467- 68 (n.a.).

[←76]

„Copil-minune” în limba germană (n.t.).